

Somatic Mutation in Cancer and Healthy Human Tissue

Sebastian Anton Friedrich Grossmann
Jesus College
University of Cambridge

September 2019

This dissertation is submitted for the degree of Doctor of Philosophy.

Summary

Somatic mutation is a natural process in healthy human tissues but has significant implications in a range of disorders and diseases with cancer being the most prominent example. In the last decade, genomics research generated big data resources that contributed tremendously to our understanding how genetic mutations and evolutionary processes shape cancer initiation and progression. More recently, sequencing efforts have shifted towards normal human tissues to explore their somatic mutational landscapes in order to understand the processes before cancer initiation. Here, data from large-scale cancer sequencing efforts is utilized as well as new data is generated to explore somatic mutation in healthy human tissues. Firstly, a mutational process that associates translocations with long-ranging epigenetic silencing is discovered in pan-cancer data. After somatic translocations involving the naturally inactivated X-chromosome in women, the RNA expression pattern of the partner chromosome indicated the spread of X-inactivation onto the autosomal partner in glioblastoma and bladder cancer. This recurrent finding highlights a novel rare mechanism how translocations contribute to genome instability and gene expression modulation in cancer. Secondly, G&T-seq, a single cell multiomics technique, is extended to single nucleotide variant analysis. When applied to cortical neurons, it suggested a substantial level of technical artefacts in previously published mutational spectra. Thus, this multiomic approach offers an alternative method in a field that is still under active development and identifies a basic but refined mutational spectrum in individual cortical neurons. Thirdly, hundreds of targeted microdissections of prostate epithelium were sequenced to investigate the lineage relationships and clonal dynamics within prostatic ductal structures. Through combination of genetic and morphological relationship between the epithelial microdissections, long-ranging developmental clones could be detected throughout the ductal structures. A secondary, more spatially confined expansion was indicated during puberty followed by tissue maintenance through local progenitor proliferation. This is the most comprehensive insight into the somatic point mutation landscape of healthy human prostate and generates valuable insight into the clonal dynamics of prostatic ductal epithelium. Collectively, this thesis contributes to the characterisation and our understanding of somatic mutational processes in health and disease.

Declaration

This thesis is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text. It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution. I further state that no substantial part of my dissertation has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution. It does not exceed the prescribed limit of 60,000 words.

Acknowledgements

The work presented in this thesis would not have been possible without the support of dedicated mentors, colleagues and collaborators.

With sincere gratitude, I would like to thank my supervisors Thierry Voet and Peter Campbell that have invested countless hours into the design of incredibly exciting projects, the critical discussion of results and conclusions as well as thoughtful advice on further analyses. While I was granted exceptional freedom, both Peter and Thierry were always present to support my progress and have made invaluable contributions to enable the presented work.

I am extremely grateful that I had the opportunity to perform my doctoral research at the Wellcome Sanger Institute, where I was able to engage with many enthusiastic and intelligent colleagues and collaborators. Sharing an office with Raheleh Rahbari, Lia Chappell, Andrew Russel, Lauren Deighton and Sabine Eckert made me look forward to come to the lab on a daily basis. Especially Raheleh provided me with helpful guidance and mentoring support. I thoroughly enjoyed working with Raheleh on complementary projects within the INSTALZ consortium. Within this consortium, I would specifically like to thank Sarah Geurs from KU Leuven for the diligent experimental work that enabled the presented work about somatic mutation in cortical neurons. Furthermore, I would like to thank Yvette Hooks, Stanley Ng, Tim Butler, Simon Brunner, Tim Coorens as well as Rakesh Heer from Newcastle University for their assistance as well as critical and insightful discussions in context of the presented work about clonal dynamics in the human prostate. While the relevant work is not presented in this thesis, I would like to thank Mia Petljak and Henry Lee-Six for involving me in their exciting projects and the fruitful collaborations. In addition to those explicitly mentioned here, several dedicated colleagues within core facilities and admin support teams contributed to this thesis and I want to express my warm gratitude to them.

Finally, I would like to thank my grandfather for having sparked my interest in science as well as my parents, sisters, friends and my partner Lisa for their continuous support.

Table of Contents

INTRODUCTION	1
Genetic Mutations	1
The Origins of Mutations and Mutational Processes	2
Germline and Somatic Mutations	3
Somatic Mosaicism in Human Tissues	4
Detection Methods for Somatic Mutations in Normal Tissues	5
Somatic Mutations in Cancer and Disease	7
Somatic Mutations in the Context of Ageing	9
Thesis Overview	10
 AUTOSOMAL SPREAD OF X-INACTIVATION AFTER SOMATIC TRANSLOCATIONS IN THE PCAWG PAN-CANCER DATA SET	 13
Introduction	13
The Role of Translocations in Human Cancer	13
X-Inactivation as Dosage Compensation Mechanism	14
Autosomal Spread of X-Inactivation	15
 Methods	 16
Data Set Description	16
Structural Variant Calls and Annotation	16
Inference of Allele-specific Expression	16
Impact of Autosome;Autosome and X;Autosome Translocations on ASE	17
Inference of X-Activity Status via Phasing	17
Inference of X-Activity Status Using Copy Number and Expression Information	18
Manual Curation of Relevant Translocations	19
Comparison of Gene Expression	20
Statistical Analysis	20
 Results	 21
Adjacency to X-chromosomal Translocations More Frequently Induces ASE Than Proximity to Autosomal Translocations	21
Activity Inference for the Translocation-involved X Chromosome Without Epigenetic Sequencing Information is Possible in a Subset of Donors	22
Somatic Translocations of the Inactive X Chromosome Can Cause Long-ranging ASE on Autosomal Partner Chromosomes	24
Autosomal Segments Affected by Long-ranging ASE Adjacent to Inactive X-Chromosomal Translocations Display Reduced Gene Expression	28

Discussion	30
INTEGRATION OF MULTIPLE DISPLACEMENT AMPLIFICATION INTO G&T-SEQ	33
Introduction	33
Single Cell Transcriptome Sequencing	33
Single Cell Genome Sequencing	35
Single Cell Multiomics	38
Methods	42
Cell Culture and Single Cell Isolation Using FACS	42
MDA Amplification and G&T-seq	42
Genotyping-by-Sequencing for MDA Product Quality Control	43
Sequencing Library Preparation and WGS Sequencing	45
WGS Quality Control and Genome Alignment	45
Computation of Depth, Breadth and Uniformity of Genome Coverage	45
Genotyping of MDA-derived Whole-Genome Sequences	46
Genome-wide SNV Calling	46
Mutational spectrum analysis	47
Statistical Analysis	47
Results	48
Experimental Setup	48
GenomiPhi Yields Insufficient Amounts of DNA	50
Pre-sequencing Quality Control Suggests Poor Amplification for Trueprime	52
Coverage Statistics are Comparable for G&T-seq and Original RepliG Products	54
Mutational Burden and Spectra are not Altered in G&T-seq RepliG	57
GbS Represents an Effective Tool for MDA Quality Control Before WGS	61
Discussion	63
A MULTIOMICS APPROACH TO DEFINE MUTATIONAL SPECTRA IN SINGLE NEURONS	69
Introduction	69
Human Cortical Development and Neurogenesis	69
Somatic Mutation in Normal Brain and in Neurodegeneration	71
Methods	75
Brain Sample Collection	75
Isolation of Single Neuronal Nuclei	75
G&T-seq for Neuronal Nuclei.	75

DNA Extraction for Bulk Sequencing Controls	76
Genotyping-by-Sequencing for MDA Product Quality Control	76
Sequencing Library Preparation, Genome and Transcriptome Sequencing	76
Genome alignments for WGS and RNA-seq Data	77
Count Matrix Generation and Quality Control for RNA-seq Data	78
Computation of FPKM Values	78
Neuronal Subtype Identification	78
SNV Calling in RNA-seq Data	79
Germline SNP Calling in Bulk WGS Data	80
Somatic SNV Calling in Single Cell WGS Data	80
Analysis of Shared Mutations from WGS Data and Phylogenies	82
Integration of SNV Calls in RNA-seq and DNA-seq Data	82
Estimation of Upper Limit for Genome-wide Burden of Mutations	83
Genome Feature Annotation	84
Association Between DNA-and-RNA SNV Calls and Gene Expression	84
Mutational Spectra and Signature Extraction	84
Gene Ontology and Pathway Enrichments	85
Statistical Analysis	85
Results	86
Fifty Neurons are Selected for the Multiomics SNV Calling Approach	86
Increased Read Depth Slightly Improves Feature Detection in RNA Data	90
Deep-sequenced RNA Confirms Excitatory and Inhibitory Neuronal Subtypes	91
Germline SNP Confirmation Rate in RNA is Limited by Bimodal VAF Distribution	94
Substantial Level of RNA Editing Detected in Single Cell RNA-seq Data	96
RNA Editing is More Prevalent in Excitatory than in Inhibitory Neurons	98
Batch Effects and Extreme Mutational Burden are Discovered in WGS Data	102
Deconvolution of Mutational Spectra and Restriction to Shared Mutations Cannot Resolve the High Artefact Background in WGS Data	107
Integrated DNA-and-RNA SNV Calls Uncover a Unique Mutational Spectrum Compared to Individual Molecular Layers	109
Integrated DNA-and-RNA SNV Calls are Enriched in Highly Expressed Genes	114
Discussion	117
MUTATIONAL PROCESSES AND CLONAL DYNAMICS IN PROSTATIC EPITHELIUM	127
Introduction	127
Anatomy and Development of the Human Prostate	127
Benign Prostate Hyperplasia and Prostate Cancer	130
Stem Cell Biology in the Normal Human Prostate	131

Methods	134
Prostate Samples	134
Sample Preparation for Laser-Capture Microscopy	134
Reconstruction of Prostatic Glandular Subunits	135
Laser-Capture Microscopy and Whole-Genome Sequencing	135
Genome Alignment and Variant Calling	136
Inference of the Relationship Between Mutation Accumulation and Aging	136
Mutational Spectra and Signature Extraction	137
Detection of Positive Selection and Driver Mutations	137
Telomere Length Estimation	138
Lineage Tree Inference	138
Statistical Analysis	139
Results	140
Data Set Description	140
Mutations Accumulate with Age in Normal Prostatic Epithelium	141
Clock-like Mutational Signatures Dominate the Mutational Landscape in Normal Prostatic Epithelium	143
No Signs of Positive Selection and Low Frequency of Driver Mutations in Normal Prostatic Epithelium	145
Long-Ranging Glandular Subunits are Reconstructed from Serial Sections of a Whole Prostate Sample	147
Increased Rates of Mutation Accumulation and Cell Division are Observed in the Peripheral Parts of Individual Glandular Subunits	151
Complex Lineage Relationships are Observed Within Glandular Subunits	153
Directed Morphogenesis is Observed During Embryonic and Pubertal Development of Glandular Subunits	159
Local Proliferation of Progenitor Cells is Suggested During Adult Tissue Maintenance	162
A Prostate Cancer Driver Mutation is Associated with Spatial Expansion of an Adult Clone	164
SBS40 is Only Detected in Recent Ancestral Clones	165
Discussion	167
CONCLUSION	175
REFERENCES	177
LIST OF ABBREVIATIONS AND ACRONYMS	197

Introduction

This thesis contributes to the characterisation of somatic mutation and mutational mechanisms in cancerous and healthy human tissues. Since the biological context in which somatic mutation is studied in differs between the individual chapters, this thesis starts with a brief summary about genetic mutations and their relevance in health and disease and a more specific introduction is presented at the start of each chapter.

Genetic Mutations

A mutation is any change in the DNA sequence of an organism. Genetic mutations appear across a large range of scales from aneuploidies over structural variations (SVs) such gross chromosomal rearrangements or focal copy number alterations down to point mutations at individual base pairs ¹. Mutations are essential to increase the genetic variability and form the basis for Darwinian evolution to act upon. The gradual and repeated selection for and removal of mutations over several generations are a main driver for speciation together with other necessary factors such as reproductive isolation ².

Since natural selection is a gradual process and the cost or conferred advantage of any mutation can vary as widely as the genomic scale affected, pronounced genetic variability can already be observed within one species. The human genome between any two individuals is identical to at least 99.5% but naturally contains high variability in the remaining fraction ³⁻⁵. While mutations that are present in more than one percent of the population are usually referred to as polymorphisms, they can be used to illustrate how common changes are in the human genome ⁶. Large-scale sequencing efforts estimate that every human carries a single nucleotide polymorphism (SNP) every 1 – 2 kilobases (kb) as well as up to over 600,000 small insertions or deletions in the range of up to 50 base pairs (bp) compared to the reference genome ⁷⁻⁹. However, not only base pair-level and small aberrations are common but each individual is estimated to carry over 2,000 larger SV events that can sometimes reach the scale of megabases (Mb) ⁹.

The Origins of Mutations and Mutational Processes

Mutations are caused by several endogenous and exogenous factors. They can arise at very low frequencies due to the molecular instability of DNA bases and spontaneous reactions ^{10,11}. Metabolic by-products such as reactive oxygen species (ROS), DNA replication before cell division as well as unwinding of the DNA double-helix during transcription all pose steady threats to DNA integrity ¹²⁻¹⁴. Additionally, cells can be exposed to environmental mutagens such as tobacco smoke and aflatoxin or UV and ionising radiation ^{15,16}. However, the accumulation of mutations is not always a gradual process as catastrophic events such as chromothripsis and kataegis can cause thousands of clustered chromosomal rearrangements or vast numbers of point mutations across a relatively short stretch of DNA ^{17,18}.

It is estimated that every single cell is facing between 70,000 – 100,000 incidences of DNA damage per day, while the relative contribution of different intrinsic and extrinsic factors is still a matter of debate ¹⁹⁻²³. To deal with this overwhelming amount of DNA damage and to maintain genomic integrity, eukaryotes have developed a DNA damage response. This response comprises several pathways that among others include base-excision repair for the correction of base modifications, nucleotide-excision repair for the correction of helix-distorting DNA lesions as well as non-homologous end-joining or homologous recombination to repair double-strand breaks ²⁴⁻²⁶. Furthermore, heavy DNA damage can also result in the induction of apoptosis to remove these genomes completely ²⁷.

Given that an adult human body consists of 10-100 trillion of cells and the daily average of DNA lesions, these repair mechanisms could face up to 10^{19} mutations per day ²⁸. Arguably, even extremely accurate repair mechanisms will fail to deal with all of these lesions correctly and will allow some mutations to manifest themselves. While the occurrence of an individual mutation is a stochastic event and can be caused by any of the factors outlined above, the type and context of mutations is not random. Each of the endogenous or exogenous factors and their associated repair mechanisms promote particular molecular reactions and thus, are associated with a spectrum of mutations. The first formalised approach to detect the association between different underlying mutational processes and their corresponding footprint was performed on the point mutation profiles of a large sequencing cohort of cancer

patients ^{29,30}. Since then, this concept has been extended using different mathematical approaches for the deconvolution of the observed spectrum into individual mutational signatures and can also consider doublet base substitutions, small insertions and deletions (indels), SVs or integrated frameworks ³¹⁻³⁷. It is important to note that for many of the mutational processes established in this way, the aetiology still remains unknown and is only linked to the independent observation of a distinct mutational signature across multiple individuals. The concept that mutations manifest themselves in a non-random context that can be linked to mutational signatures will be used in several analyses in this thesis.

Germline and Somatic Mutations

In multicellular organisms, it is important to distinguish between germinal and somatic mutations. Any germline mutation has the chance to be passed on to the progeny and will be present in all cells of the newborn. The germline mutation rate must correspond to the rate of natural selection as a too high mutation rate can cause accumulating deleterious effects while excessive fidelity is a fitness cost in itself and prevents adaptability to environmental changes. Therefore, the mutation rate can vary widely between species ³⁸⁻⁴¹. The germline point mutation rate in humans was estimated to be about $1-1.2 \times 10^{-8}$ per nucleotide per generation using trio-sequencing approaches, which corresponds to roughly 60 new mutations per generation ^{42,43}.

In contrast to germline mutations, somatic mutations are not necessarily required for evolution since they cannot be passed on to the next generation. The evolutionary argument for their existence is that excessive fidelity in the soma would represent a fitness cost with little added benefit past the reproductive age ^{39,41,44}. Although somatic mutations cannot be passed on through sexual reproduction, any somatic mutation occurring in a dividing cell will be present in all of its cellular progeny. This corresponding population of cells is called a clone and the parallel existence of many of such clones within one individual was termed somatic mosaicism ^{45,46}. Since the concept of somatic mosaicism is central to this thesis, I will give a brief perspective about its relevance in the human body and the impact on earlier studies exploring somatic mutation.

Somatic Mosaicism in Human Tissues

While early studies already indicated that the somatic mutation rate is orders of magnitude higher than in the germline, the initial efforts were confined to observations at reporter loci for technical reasons ⁴⁷⁻⁴⁹. Several years later, the advent of next-generation sequencing technologies allowed for a genome-wide detection of mutations ⁵⁰. However, as the DNA of up to one million cells is used as input material for classic bulk genome sequencing approaches, mutations can only be found if their clonality exceeds the detection limit ⁵¹.

Clone sizes depend on the proliferative capacity of the cell a mutation originally occurs in. Thus, the corresponding clone sector is usually larger the earlier the mutation is acquired in development ⁵². Several studies have explored developmental somatic mutations in monozygotic twins and successfully challenged the view that any phenotypic difference in these twins has to be due to environmental factors ⁵³⁻⁵⁹. Another study used the relative frequencies of blood cells carrying different developmental mutations and could show that many early embryonic cell doubling events contribute to the adult blood in an asymmetric 2:1 ratio, further increasing the evidence of how pronounced developmental somatic mosaicism is in humans ⁶⁰.

While the most pronounced somatic mosaicism is established during development, postnatal tissue maintenance and environmental exposure steadily increases the genetic variability and can give rise to new clones ⁶¹. However, somatic mutations acquired in this period are usually only present in few cells and lack sufficient clonality to be detected by classic bulk sequencing approaches ⁵¹. To circumvent this issue, initial efforts focussed on clonal haematopoiesis and cancer as a large number of cells deriving from a single or few somatic progenitors can be sampled in these instances ⁶²⁻⁶⁶. The studies focussed on clonal haematopoiesis could demonstrate that clonal detectability of mutations becomes more common as the age of the donor increases. Although the risk for cancer increases with this observed clonal haematopoiesis, most of the mutations were not predicted to have any functional impact ⁶²⁻⁶⁴. An even more striking association with age and a first quantitative estimate of the inter-tissue variability of somatic mutation rates and spectra was derived by investigating mutational signatures in a large cancer sequencing cohort of various tissue origins. In nearly all tissues, two clock-like mutational signatures could be demonstrated.

Remarkably, the total and relative burden introduced by these processes varied greatly between tissues. The authors suggested that one signature was due to mitotic division while the other signature was the footprint of an unrelated clock-like process ⁶⁶.

In summary, somatic mosaicism is continuously shaped and increases during an individual's developmental and postnatal lifetime. Initial efforts seeking to characterise somatic mutations on sub-microscopic detail were restricted to developmental mutations or clonal syndromes and diseases to circumvent technical detection limits.

Detection Methods for Somatic Mutations in Normal Tissues

Several approaches have been developed to overcome heterogeneity in normal human tissues and the technical detection limit of classic bulk genome sequencing (reviewed in ^{51,67}). These approaches can be divided into error-corrected ultra-deep sequencing, *in-vitro* clonal culture or (oligo-)clonal microbiopsies followed by regular sequencing or single cell sequencing.

Since the detection limit of genome sequencing data is a function of the obtained read depth, ultra-deep sequencing can uncover normal tissue heterogeneity to a certain extent. Estimates of variant allele fractions (VAFs) can be used to estimate clone sizes and thus, the clonal architecture of patches of tissues could be reconstructed in skin and oesophagus ⁶⁸⁻⁷⁰. However, deep sequencing approaches are eventually limited by the rate of technical artefacts during library preparation and sequencing. To improve the detection limits, several DNA barcoding approaches have been developed. The barcode introduction allows for the assembly of read families that originate from the same DNA molecule or even from the corresponding Watson or Crick strands. Stochastic PCR and sequencing errors will only be present in part of the read family, which allows for more reliable variant calls ⁷¹⁻⁷³. Further enrichment of these read families by rolling replication of the barcoded DNA molecules or a limiting dilution step before library amplification finally allow for variant detection at frequencies as low as $10^{-7} - 10^{-9}$ mutations per base pair ^{74,75}. For economic and technical reasons, these approaches are limited to targeted or randomly sampled regions of the genome and thus, do not allow an unbiased genome-wide investigation of somatic mutations.

Clonal culture approaches reduce the detection-limiting heterogeneity within normal tissues by culturing single cell-derived clones. Since all somatic mutations of the founder cell are now clonal, their mutational landscape can reliably be revealed using classic bulk genome sequencing. Since sufficient clonal culture sizes are difficult or impossible to obtain for several terminally differentiated human tissues, several studies have used reprogramming approaches to derive induced-pluripotent stem cells (iPSCs) before clonal expansion. However, the creation of iPSCs is inherently mutagenic ⁷⁶⁻⁷⁹. This mutagenic process can be circumvented by somatic nuclear cell transfer (SNCT) into enucleated oocytes. This cloning approach depends on the epigenetic plasticity of the transferred nucleus and was criticized to preferentially work in cells carrying favourable mutations, which would result in a biased estimate of mutational burden and processes ^{80,81}. A more direct approach is to utilize tissue-resident stem or progenitor cells to proliferate in culture. While this process can be very inefficient for some tissues, clonal patches or organoids can be derived for a wide range of tissues including stomach, small and large bowel, prostate, skin, liver, myeloid and lymphoid lineages and even neurons ⁸²⁻⁸⁷. However, even these approaches might suffer from an inherent selection bias and some of the cell culture-induced mutations are difficult to distinguish from the original *in-vivo* acquired somatic mutations.

Genetic heterogeneity can also be reduced by using targeted microdissections as input for bulk sequencing. Laser-capture microdissection (LCM) is a powerful technology to restrict the input material to a morphological subunit or a defined cell type within a heterogeneous tissue ⁸⁸. With sufficient morphological constraints, it is possible to derive samples that share a very recent common ancestor and thus, are clonal for all mutations present at that time. Colonic crypts are a particularly well-suited example to illustrate this since all cells in individual crypts are derived from a single progenitor ^{89,90}. Recent studies have used this technology to study somatic mutation in colorectal and endometrial epithelial cells ⁹¹⁻⁹³. In this thesis, all insights into somatic mutation in the human prostate and clonal dynamics within individual ductal trees are derived on the basis of targeted microdissections using LCM.

The most obvious way to circumvent genetic variability within a sample is to sequence the DNA of single cells. In theory, this approach is suitable to study any cell type, is

not dependent on tissue architecture and cells can be selected at random to avoid any bias towards epigenetic or mutational status. Several studies have explored different classes of somatic mutations, including single nucleotide variants (SNVs), retrotranspositions and copy-number variations (CNVs) in skin, heart and predominantly in brain ⁹⁴⁻¹⁰⁴. However, genome sequencing of single cells requires whole-genome amplification (WGA) and can result in uneven and incomplete genome coverage, substantial error rates and thus, suboptimal sensitivity and specificity ^{105,106}. As single cell sequencing is used to characterise the somatic SNV spectrum in human neurons in this thesis, a more detailed discussion of single cell sequencing and the differences in sensitivity for different types of mutations follows in the relevant chapters.

Using the methods explained in the previous section, several studies have recently unveiled surprising findings about somatic mutations in healthy human cells while highlighting both common features as well as great differences between tissues (summarised in ^{67,107}). Most of these studies were motivated by cancer research and to fully appreciate the findings in normal tissues, understanding of the relevance of somatic mutation in disease is necessary.

Somatic Mutations in Cancer and Disease

DNA mutations are essential for evolution and genetic variation is common in healthy tissues. However, particular germline and somatic mutations are linked to several disorders and diseases. A plethora of syndromes, diseases and disorders are associated with inherited and *de-novo* germline mutations. While they are highly relevant to human health, this thesis contributes to the research of somatic mutations. As outlined above, methods to study germline and somatic variation differ as do their implications. Therefore, this section only aims to give a brief summary about the history and relevance of somatic mutations in disease.

Cancer is the most well-characterised example of a disease that is caused by somatic mutation ¹⁰⁸. The relationship between mutations and cancer emerged in the late nineteenth and early twentieth century, when the German biologists David von Hansemann and Theodor Boveri observed abnormalities in mitotic divisions and corresponding chromosomal aberrations in cancerous epithelial cells ^{109,110}.

Extensive research over several decades identified several chemical and environmental factors that were epidemiologically associated with cancer and could induce neoplasms in animal models. After the discovery of DNA as the inheritable substance and its structure in the forties and fifties, these mutagens were shown to cause chemical changes in the DNA, further strengthening the link between genetic alterations and the occurrence of cancer (reviewed in ¹⁵).

Refined chromosomal analyses in the sixties and seventies could demonstrate the association between a recurrent chromosomal translocation and a particular cancer type. The translocation between chromosomes 9 and 22 and the correspondingly derived Philadelphia chromosome is a hallmark of chronic myeloid leukaemia (CML) that can also be present in rare subtypes of other leukaemias ¹¹¹⁻¹¹³. Although the concept was only defined much later, this recurrent structural variation was the first identification of a driver mutation. Driver mutations are deemed to be the driving force for cancer initiation and progression as they confer a selective growth advantage to its carrier. Contrary to this, so-called passenger mutations have either negligible minor or neutral effects ^{108,114}. It is important to note that the vast majority of mutations detected in cancer are passenger mutations. On average, a cancer genome is estimated to have four driver mutations with substantial variation depending on the tissue type of origin ¹¹⁵⁻¹¹⁷.

Without a precise definition of driver and passenger mutations, several important cancer genes and concepts were discovered in the seventies and eighties. By comparing the age-incidence-relationship of familial and sporadic retinoblastoma, the requirement for two independent hits in particular genes was postulated for cancer progression and the corresponding RB1 gene was identified as the first tumour suppressor gene ^{118,119}. Additionally, oncogenes were discovered through the neoplastic potential of viral oncogene homologs ^{120,121}. Finally, it could be shown that the introduction of genomic DNA from cancerous cells into phenotypically normal cell lines caused their transformation, which was the final proof for the causal relationship of genetic mutations and cancer ^{122,123}.

In the following decades until today, a myriad of individual studies as well as several large-scale consortium projects such as the International Cancer Genome Consortium

(ICGC), The Cancer Genome Atlas (TCGA) and the Pan-Cancer Analysis of Whole Genomes Project (PCAWG) have contributed to our extensive knowledge about genetic mutations in cancer ¹²⁴⁻¹²⁶. Currently, 719 genes are included in the Cancer Gene Census and COSMIC, the largest cancer-related database, describes nearly 6 million coding mutations across 1.4 million cancer samples that have been curated from a total of over 26,000 publications ¹²⁷.

The most extensive research of somatic mutation as a cause for disease was performed in the context of cancer but several other examples have been identified. Since only a subset of cells within one organism is affected by any particular somatic mutation, somatic mosaicism can result in attenuated phenotypes of well-established Mendelian diseases. Moreover, some diseases can only manifest in this mosaic state when the corresponding constitutional mutation is lethal during development ¹²⁸. Especially developmentally acquired somatic mutations were shown to be associated with a range of diseases including psychiatric disorders, autoimmune and cardiovascular diseases as well as complex syndromes ^{55,58,107,128-132}. In recent years, ample evidence has also been found for the relevance of somatic mutations in neurodegeneration and this is briefly summarised in the chapter addressing somatic base substitutions in single human neurons.

Somatic Mutations in the Context of Ageing

Somatic mutations are more frequently occurring during development but also manifest later in life and over the course of aging ⁶¹. Several recent studies could demonstrate this age-dependent increase of somatic mutation directly in normal human tissues while it was already implicated by their corresponding cancerous genomes as previously explained ^{66,68,70,84,86,91,92,102}. This accumulation of mutations and the functional decline of tissues and organs with ageing, has previously been associated with each other in the mutation theory of aging ^{133,134}. Originally, this model was criticised as the somatic burden in normal cells was not thought to be substantial enough to cause functional decline of a whole tissue ^{85,107,135}. However, the recent genomic surveys of normal tissues have revolutionised our understanding of their mutational landscape and the resulting clonal architecture. Sun-exposed skin cells were shown to carry more than 10,000 mutations in total and a positively selected mutation in known cancer genes was present in every fourth cell ⁶⁹. While the total

mutational burden was found to be much lower in normal epithelium of the oesophagus, an even higher prevalence of known driver genes and convoluted clonal architecture was discovered ^{68,70}. However, there is also great disparity between tissues as the high prevalence of driver mutations in skin, oesophagus and endometrium is contrasted with their near absence in normal blood and colon ^{86,91-93}.

The universal presence of selection in normal tissues and the complex clonal architecture discovered within them, yields new insight into the relationship between somatic mutations and the functional decline with age ^{68-70,86,91-93,107,117}. While it is unlikely that enough normal cells acquire independent mutations that lead to a progressive loss of function, positively selected clones can overtake larger parts of the tissue. In essence, this reduces clonality and diversity within a tissue and is similar in nature to what had already been observed in clonal haematopoiesis ^{62-64,85}. This less diverse tissue architecture could contribute to the progressively declining function of somatic tissues.

Thesis Overview

On the one hand, this thesis leverages large pan-cancer data sets to explore somatic mutational mechanisms. On the other hand, it also contributes to and applies recent methods to explore somatic mutation in normal human tissues.

The presented work is divided into four main chapters. The first chapter utilizes the PCAWG pan-cancer data set to uncover a mutational mechanism that links somatic translocations to long-ranging epigenetic silencing on the translocated chromosomes. The second chapter demonstrates that a method for the characterisation of point mutations in single cells can be used within a multi-omics technique that allows parallel genome and transcriptome sequencing of single cells. The third chapter applies this technique to characterise the mutational burden and spectrum in individual neurons from normal human brains. The fourth chapter explores somatic mutations within the healthy human prostate and provides a qualitative assessment of the clonal dynamics within individual ductal networks.

Autosomal Spread of X-Inactivation After Somatic Translocations in the PCAWG Pan-cancer Data Set

Introduction

This chapter describes the spread of epigenetic silencing from the inactivated X chromosome to autosomes involved in somatic translocations in the Pan-Cancer Analysis of Whole-Genomes (PCAWG) data set. Two examples from different cancer types are identified that show overall decreased and allele-specific expression (ASE) of whole chromosome arms from translocation partners of the inactive X chromosome. Due to the resemblance to X chromosomal expression patterns and in absence of any further structural variations, these examples are suggested to represent autosomal spread of X-inactivation.

The importance and role of translocations in human cancer as well as X-inactivation and the known cases of autosomal spread are briefly summarised in this chapter.

The Role of Translocations in Human Cancer

Genome instability and mutation is one of the hallmarks of cancer ¹³⁶. While most classes of mutations also occur in normal somatic tissue as explained in the general introduction, the acquisition of structural driver events is suggested to be a rate-limiting step in tumorigenesis ¹³⁷. Notably, the Philadelphia chromosome as consequence of a recurrent translocation between chromosome 9 and 22 was the first discovered cancer driver gene ^{111,112}. In general, translocations have a crucial role in cancer initiation and progression and are estimated to drive about 20% of all cancer cases ¹³⁸.

Traditionally, translocations are classified as unbalanced (associated with large-scale copy number change) or balanced (copy number-neutral) and distinguished between intra- and inter-chromosomal translocation involving two loci of the same or of two distinct chromosomes ¹³⁹. Unbalanced translocations typically exert their oncogenic effects through gene expression dosage consequences of the copy number changes ¹¹⁶. Balanced translocations can generate chimeric fusion genes or affect gene expression through juxtaposition with promoters or other regulatory

elements^{138,140}. Notably, most cancer-relevant translocations are thought to increase expression of deregulated genes and the vast majority of fusion genes are associated with oncogenic function^{138,140}. However, rare examples for a suppressive function on gene expression or activity such as TTC28 fusions in some colorectal cancer patients exist¹⁴¹.

While translocations can be caused by mechanisms like chromothripsis and chromoplexy or be part of other complex structural variations, traditionally defined translocations are caused by the erroneous repair of two simultaneously occurring double-strand breaks (DSBs)^{17,35,140,142}. While the exact process is not fully resolved yet, it has been shown that several epigenetic modifications are involved in this repair process, highlighting the importance of dynamic chromatin modification for the generation of translocations¹⁴³. While these local changes around the breakpoints are dynamic and temporary, the resulting translocation has the potential to join genomic regions with different chromatin states. Chromatin states can largely be divided into condensed and transcriptionally inactive heterochromatin as well as lightly packed and accessible euchromatin¹⁴⁴. A multitude of epigenetic mechanisms influence the chromatin state including binding of non-coding RNAs, DNA methylation, histone modification and nucleosome positioning¹⁴⁵. Several of these marks are mediated by DNA-bound chromatin remodelers that create local gradients around their own position and can facilitate spreading by recruiting similar chromatin remodelers in a positive feedback loop^{145,146}. Therefore, translocations between heterochromatic and euchromatic regions of the genome have the potential for epigenetic dysregulation of the translocation partner.

X-Inactivation as Dosage Compensation Mechanism

Different dosage compensation mechanisms have evolved in the animal kingdom to account for the different amount of gonosomes between the sexes of the same species¹⁴⁷. In humans, one of the two X chromosomes in females is randomly selected for X chromosome inactivation (XCI) in the late blastocyst stage that causes life-long epigenetic silencing of the selected copy in all somatic cells^{148,149}. For the initiation of XCI, the expression of the long non-coding RNA XIST from the cis-acting Xist locus of the inactive X chromosome plays a central role^{149,150}. During the XCI process, Xist progressively coats the whole inactive X chromosome and a facultative

heterochromatic state is promoted by several additional chromatin modifications, including repressive histone tail modifications, DNA hypermethylation and nucleosome positioning^{151,152}. While about 15% of X-chromosomal genes consistently escape the epigenetic silencing to a certain extent and about 10% display a variable pattern of inactivation, the facultative heterochromatin is maintained throughout life and re-established on the same copy in progenitor cells^{147,151,153}. Notably, XCI initiation and maintenance are at least partially dependent on different mechanisms since deletion of Xist does not interfere with XCI maintenance after the initial establishment^{152,154}.

Autosomal Spread of X-Inactivation

While only additional copies of the X chromosome are subject to physiological chromosome-wide silencing, spreading of the heterochromatic features and concomitant gene silencing has been shown for several cases with germline X;autosome translocations¹⁵⁵⁻¹⁶¹. For all of these cases, the autosome silencing is thought to arise from spreading of physiological X chromosome inactivation during development and subsequent maintenance. While more genes on affected autosomes seem to escape the complete silencing compared to the inactive X chromosome, expression levels of whole chromosome arms are affected and were associated with developmental disorder or increased cancer risk^{156-158,160-162}.

While these studies highlight the potential for autosomal spread of XCI for germline translocations before the physiological establishment of the inactive X chromosome, only one case study describes the autosomal spread of XCI after a somatic translocation. In this case, gene silencing of chromosome 5 close after a somatically acquired X;5 translocation could be identified as one of the driver mutations for essential thrombocythemia¹⁶³. Given the differences for developmental establishment and adult maintenance of XCI, corresponding differences in the efficiency of autosomal spread of XCI after somatic translocation compared to germline events are reasonable. Here, the frequency of this largely unexplored mutational mechanism is considered in a pan-cancer data set comprising whole-genome and transcriptome data from 620 female patients.

Methods

Data Set Description

The PCAWG data set is currently the largest collection of whole-genome sequencing data for cancer research with a total of 2,834 donors from 39 distinct tumor types ¹²⁶. Additionally, transcriptome data is available for 1,188 from 27 cancer types ¹⁶⁴.

Considered samples were restricted to female patients with whole-genome and whole-transcriptome data that passed the PCAWG consortium quality control (QC) standards ¹⁶⁵. For genome sequencing, main QC criteria assessed were the mean depth and the evenness of coverage, the ratio of paired reads mapping to different chromosomes, and the ratio of substitutions between forward and reverse reads. For RNA-seq, the QC metrics included sequence quality scores such as base-wise quality scores and GC content, mapping quality scores including a minimum fraction of 50% mapped reads and an assessment of the 3'/5' bias. While transcriptome data is available as Tophat2 and STAR alignments, only the STAR alignments were considered for the presented analyses. In total, data for 620 patients from 26 cancer subtypes was obtained.

Structural Variant Calls and Annotation

Structural variant calls from the BRASS algorithm generated by the PCAWG Structural Variation Working Group were downloaded for the same female cancer patients described above ^{35,166}. The downloaded SV calls included annotation that distinguished classes of structural variants such as deletions, inversions or translocations. Only structural variants annotated as translocations were considered.

Inference of Allele-specific Expression

Unphased germline variants for the same female cancer patients were downloaded as generated by the PCAWG Germline Working Group ¹⁶⁷. Allele-specific counts for these variants were generated for the genome and transcriptome data. For allele-specific counts, only bases with phred-scaled quality score of at least 20 on reads with a phred-scaled mapping quality of at least 20 were considered. Subsequently, computation of ASE was restricted to loci with at least eight counts each in WGS and

RNA-seq data. Both WGS and RNA-seq data were considered to account for potential DNA copy number related changes in expression.

The per-variant odds ratio and significance for allele-specific expression were computed using Fisher's exact test. To analyse allele-specific expression on a gene level, the significance for all variants within protein-coding genes according to the GENCODE 19 annotation was combined using Fisher's method and corrected for multiple testing using the Benjamini-Hochberg procedure.

Impact of Autosome;Autosome and X;Autosome Translocations on ASE

All translocations can interfere with physiological gene expression of their new genomic environment. Thus, potentially observed ASE is not necessarily due to autosomal spread of XCI and a uniformly distributed test statistic is not expected when testing genes for ASE in adjacency to X;autosome translocations.

To reveal differences with respect to ASE in vicinity to autosome;autosome or X;autosome translocations, ASE was computed for all autosomal protein-coding genes as defined in the GENCODE 19 annotation within a maximum distance of 4 Mb from annotated breakpoints of heterologous chromosome translocations. In case one gene was within the distance and correct orientation of several translocation breakpoints, the gene was associated with the closest translocation. A bootstrap hypothesis test of equality was performed to compare the p-value distributions for autosomal genes in vicinity of X-chromosomal or autosomal translocations using the R package `sm` ¹⁶⁸.

Inference of X-Activity Status via Phasing

All reads supporting the annotated X-chromosomal translocations were screened for the presence of heterozygous germline variants from the respective sample. In case one SNP allele could be associated with a translocation breakpoint, the X chromosome was phased for this sample using SHAPEIT v2.r837 against the 1000 genomes reference panel ¹⁶⁹. To improve the quality and reduce branching of the haplotype tree structure, the read-aware phasing option was used ¹⁷⁰. Phasing uncertainty of the breakpoint-supported variant with the three closest expressed heterozygous SNPs both upstream and downstream was computed using

SHAPEIT v2.r837 ¹⁶⁹. A total of 1,000 discrete haplotype sets were sampled from the phasing graph structure emitted by SHAPEIT v2.r837 and the probability for all haplotype combinations of the selected germline variants was calculated. Phasing with the breakpoint-supporting variant was considered sufficient if a haplotype block of at least three of the seven considered variants displayed a phasing likelihood greater than 0.9. For these cases, the X-activity status was derived by comparing the haplotype for the breakpoint-associated allele with the haplotype expression status and assigned accordingly.

Inference of X-Activity Status Using Copy Number and Expression Information

For some translocations, the activity status of the X chromosome could be inferred using copy number information. Copy number analyses were generated by the PCAWG Structural Variation Working Group using ascatNgs and downloaded for the patients considered in this study ^{35,166}.

By definition, all unbalanced translocations coincide with a change in copy number. However, also reciprocal translocations frequently lead to a small deletion in at least one of the chromosomes involved in the translocation ¹⁷¹. For two different scenarios, ploidy information and copy number changes were used to assign the activity status to the X chromosome involved in an X;autosome translocation.

In case of a single copy of the X chromosome, samples were checked for X-chromosomal ASE. In case the single X chromosome copy in the tumour would be silenced, any X-chromosomal expression must have derived from the non-clonal normal contamination in the tumour sample. In this case, most loci would show expression of both X-chromosomal alleles. However, if substantial ASE was observed across the whole X chromosome, the single X-chromosomal copy was actively expressed in the tumour sample. Therefore, all X-chromosomal translocation in this sample were assigned to involve the active X chromosome. As ASE was always maintained for samples with a single copy of the X chromosome, only active X;autosome translocations were identified for these cases.

In case two X-chromosomal copies were observed, large coinciding deletions were necessary to assign the activity status to the X chromosome involved in the

translocation. Notably, it was necessary to assume the most parsimonious explanation of nearly identical breakpoints for deletions and translocation, in which the deletion and the translocation involve the same DNA molecule and not both homologous chromosomes. Similar to the example above, ASE of the X chromosome across the deleted regions was used to assign the X-activity status. If ASE was maintained for the monosomic segment, the inactive copy must have been lost and thus, been the one involved in the translocation. Accordingly, if ASE was lost for the monosomic segment, the active X chromosome must have been involved in the translocation. Notably, only X;autosome translocations involving the inactive copy were detected using this method.

Manual Curation of Relevant Translocations

All breakpoints annotated as X;autosome translocation with assigned X-activity status were manually curated. All structural variants and copy numbers of the X chromosome and the autosomal translocation partner were visualised as described by the PCAWG Structural Variation Working Group ³⁵. Any translocations similar to templated insertions, chromothripsis, chromoplexy or other complex structural variation on the involved chromosome arms were excluded from further analyses. Templated insertions were characterised by a focal copy number gain of the donor chromosome, a focal deletion of the receiving chromosome and the corresponding orientation of two inter-chromosomal breakpoints ³⁵. Chromothripsis was defined as high translocation breakpoints to various chromosomes. Chromoplexy was assumed if at least three interlinked translocations were observed. Any translocation involving more than two chromosomes or coinciding deletion or duplication was considered to be a complex event and also excluded.

Comparison of Gene Expression

Autosomal spread after translocations involving the inactive X-chromosome was indicated in a glioblastoma (GBM) and bladder urothelial cancer (BLCA). Gene expression information for 28 additional GBM and 21 BLCA patients were downloaded as FPKM values as generated by the PCAWG Transcriptome Working Group ¹⁶⁴. All genes with FPKM values of at least 0.1 in the sample with X;autosome translocation were considered as expressed and the cancer-specific mean expression for these genes was computed using the additional samples from the corresponding cancer subtype.

Statistical Analysis

All statistical analysis was performed in R version 3.5.0 using core distribution functions unless otherwise indicated ¹⁷².

Results

Adjacency to X-chromosomal Translocations More Frequently Induces ASE Than Proximity to Autosomal Translocations

Physiological XCI in women is necessary for dosage compensation and results in ASE for most X-chromosomal genes. Since XCI is initially randomly established per cell in the human blastocyst and then clonally maintained throughout life, normal tissues do not display ASE for X-chromosomal expression in bulk RNA-seq¹⁴⁹. However, all cells within one tumour originally derive from a single somatic cell and thus, display repression of the same X chromosome. Since no epigenetic information was available for most donors from PCAWG, ASE was used as one proxy for XCI.

All structural variants, including autosome;autosome translocations, can cause ASE in their adjacency by disrupting gene boundaries or regulatory elements. To ensure that ASE marks the autosomal spread of XCI above this random background, I compared the frequency of autosomal genes displaying ASE adjacent to X-chromosomal or autosomal translocations as described in the methods section.

In the 620 appropriate samples from PCAWG, we identified a total of 85,584 inter-chromosomal breakpoints, with 2,077 (2.4%) from 158 donors being X;autosome inter-chromosomal breakpoints. ASE was evaluated for 8,606 and 229,087 autosomal genes in proximity to X;autosome and autosome;autosome inter-chromosomal breakpoints, respectively. After correction for multiple hypothesis testing using the Benjamini-Hochberg procedure, a total of 1,574 (17.9%) and 35,124 (15.3%) autosomal genes adjacent to X;autosome and autosome;autosome inter-chromosomal breakpoints, respectively, displayed significant ASE (defined by $q < 0.05$). While significant ASE was also detected near autosomal translocation breakpoints, genes associated with X-chromosomal translocation breakpoints were more significantly affected (see Figure1; bootstrap hypothesis test of equality, $p < 1e-3$). This indicated that ASE is more common in proximity of X-chromosomal translocations compared to a general translocation background and suggested a unique effect of the X chromosome.

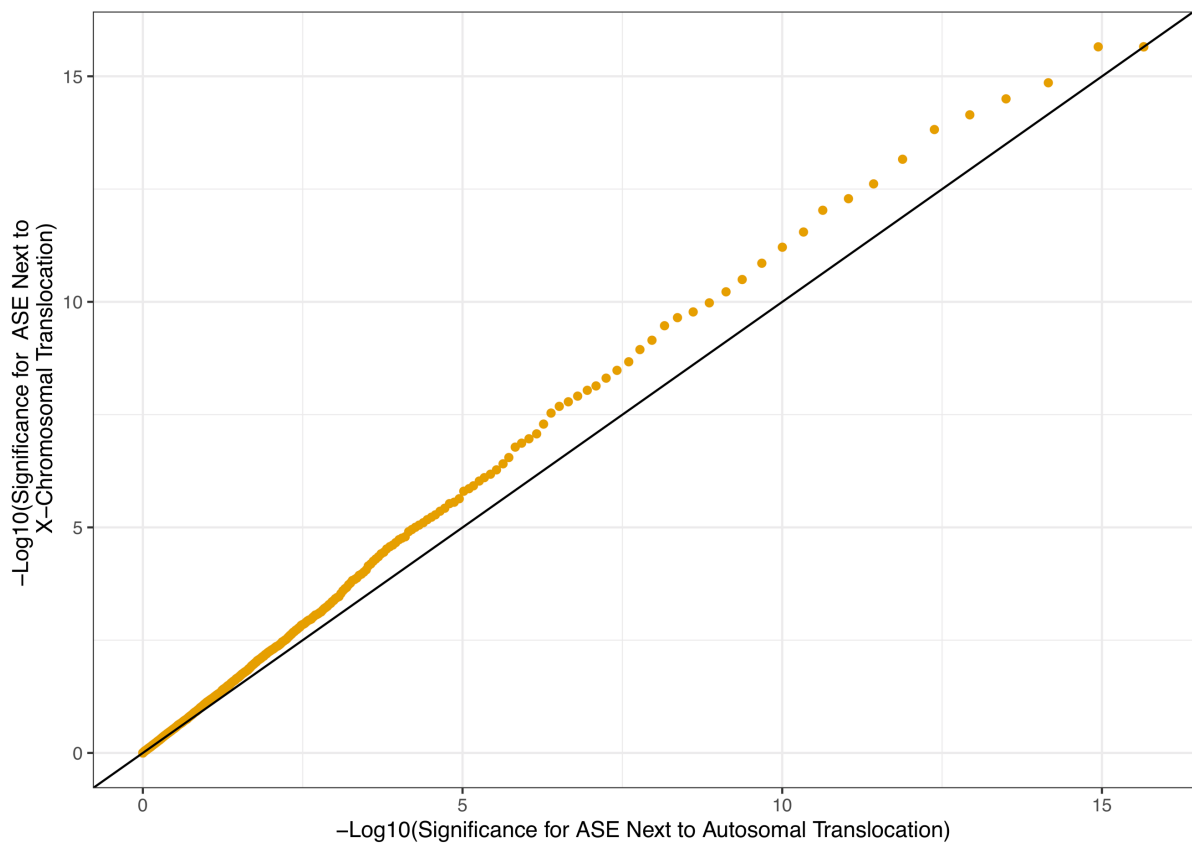


Figure 1: Autosomal Genes Near X-Chromosomal Translocations are More Significantly Affected by ASE than Genes Near Autosomal Translocations. The probability for ASE was calculated for autosomal protein coding genes and associated with translocation breakpoints as called by BRASS. Genes were associated with the closest translocations within a 4 MB window and annotated as being affected by either an autosome;autosome or an X;autosome translocation. The QQ-plot demonstrates that autosomal genes in proximity to an X-chromosomal translocation are more significantly affected by ASE compared to a background defined by autosomal translocations.

Activity Inference for the Translocation-involved X Chromosome Without Epigenetic Sequencing Information is Possible in a Subset of Donors

To further explore this apparent bias in skewed allele expression, it was important to establish whether the inactive or the active X chromosome was involved in the breakpoint. Since there was no epigenetic information available for any of the 158 samples with X;autosome translocations, three different approaches relying on copy number, expression and phasing data were used to infer the X-activity status, which are described in the methods section in more detail: i) In cases where only a single copy of the X chromosome was present, all translocations in this sample could be assigned to involve the active copy (Figure 2A); ii) If a translocation breakpoint on the

X chromosome coincided with a copy number loss from two to one copy of the X chromosome and ASE was maintained in the lost segment, the inactive copy was implied to be involved in the translocation (Figure 2B); iii) If at least one of the reads supporting the breakpoint on the X chromosome featured a heterozygous germline variant, haplotype-phasing with other expressed variants on the X chromosome was performed and the activity status of the X chromosome involved in the translocation assigned accordingly if high phasing quality was indicated. In total, the X-activity status could be assigned for 61 X;autosome breakpoints from 23 samples. Of these 61 translocation breakpoints, 40 breakpoints in 9 samples were assigned to involve the active X chromosome, with the remaining 21 breakpoints in 14 samples were associated with translocations of the inactive X chromosome.

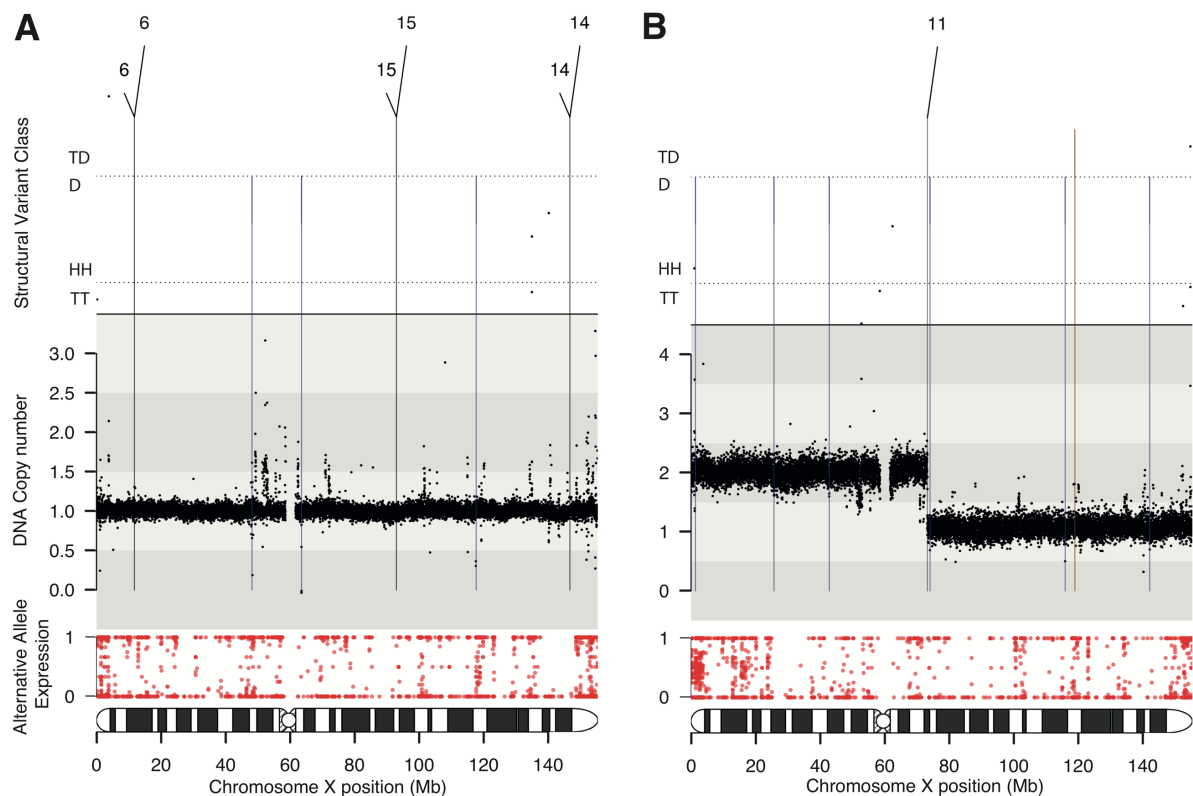


Figure 2: Copy Number and Expression Information Can be Used to Infer the X-Activity Status. TD: Tandem duplication. D: Deletion. HH: Head-to-head inversion. TT: Tail-to-tail inversion. **(A)** A single copy of the X chromosome with chromosome-wide ASE indicates the involvement of the active X chromosome in the three X;autosome translocations in this sample. **(B)** Large scale deletions resulting in loss of one of the two X-chromosome copies that coincide with an autosomal translocation characterize unbalanced X;autosome translocations. ASE in the monoallelic region indicates that the inactive copy was lost. The 11;X translocation in this sample involved the inactive X chromosome.

Subsequent to X-activity inference, the subset of translocation breakpoints with assigned status was further curated to exclude all translocation-annotated variants that did not represent single large-scale translocations. Two breakpoints within one sample were interpreted to derive from chromothripsis, which would obscure any inference about autosomal spread of X-silencing. Moreover, a total of 37 breakpoints from 16 samples represented templated insertions. Templated insertions are a common SV in cancer, where short chromosomal segments are copied and inserted into distant regions of the genome, presumably during replication³⁵. Since these short replication-based events are unlikely to confer any spread of their epigenetic status, they were also excluded from further analyses.

The remaining 22 X;autosome breakpoints with assigned X-activity status could be confirmed to represent genuine translocation events. Considering that one balanced translocation is marked by two and unbalanced translocations by one X-chromosomal breakpoint, a total of 17 large-scale translocations with inferred X-activity status could be identified. From the total of 17 curated X;autosome translocations, eleven events in six donors were identified to involve the active copy and six translocations from another six donors were associated with the inactive copy of the X chromosome.

Somatic Translocations of the Inactive X Chromosome Can Cause Long-ranging ASE on Autosomal Partner Chromosomes

Autosomal spread of XCI after germline X;autosome translocations was previously shown to be long-ranging but discontinuous¹⁶². To assess ASE potentially caused by autosomal spread of XCI in the presented data set, the odds ratio for ASE was computed for all autosomes involved in the 17 manually curated X;autosome translocations with inferred activity status of the X chromosome.

In the six donors that featured the 17 events with active X translocations, the corresponding autosomes only displayed ASE on chromosome segments that were affected by additional structural variations such as copy number loss or loss of heterozygosity (LOH) (Figure 3). Hence, no substantially elevated odds ratio for ASE on translocation partners of the active X-chromosome could be observed.

The six translocations involving the inactive X chromosome were detected in three patients with ovarian serous carcinoma (OSC) as well as in one hepatocellular carcinoma (HCC), GBM and BLCA. The three OSC as well as the one HCC case did not display substantially elevated odds ratios for ASE on the autosomal part of the chromosome that was adjacent to the inactive X chromosome following the translocation (Figure 4).

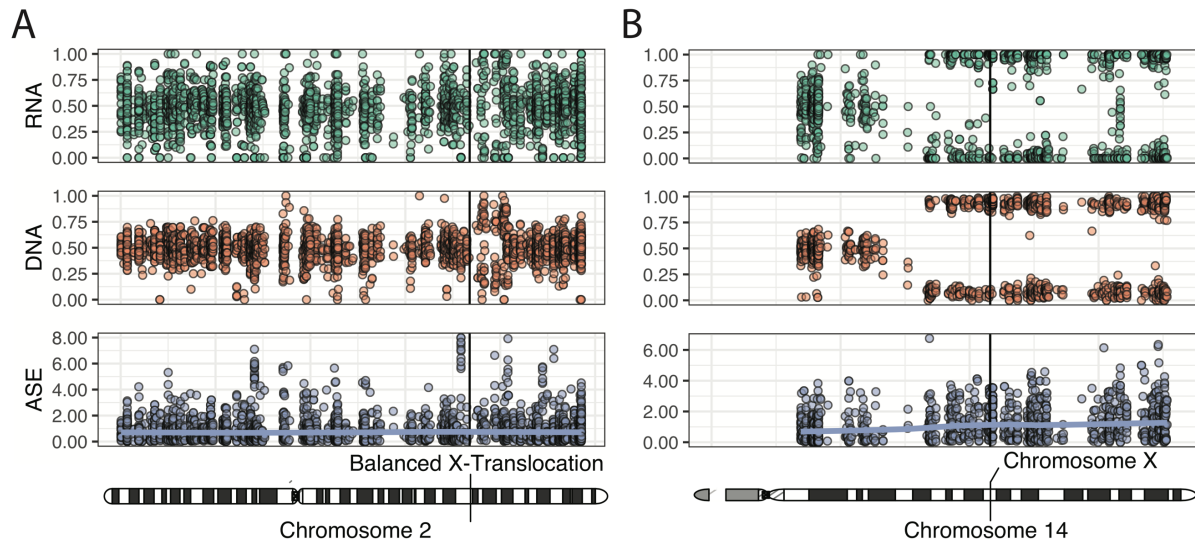


Figure 3: Translocations of the Active X Chromosome Do Not Induce ASE. The alternative allele fraction detected in RNA-seq and WGS as well as the corresponding log odds ratio for ASE were computed for all heterozygous germline SNPs on the autosomal partner chromosome of active X-chromosomal translocations. The average log odds ratio is indicated by the purple spline fit and the breakpoint of X-chromosomal translocations is highlighted in black. **(A)** No elevated ASE is detected on either side of chromosome 2 after a balanced translocation involving the active X chromosome in a malignant lymphoma sample. **(B)** ASE observed on chromosome 14 after an unbalanced X-chromosomal translocation is associated with LOH in a breast cancer sample and does not imply autosomal spread of XCI.

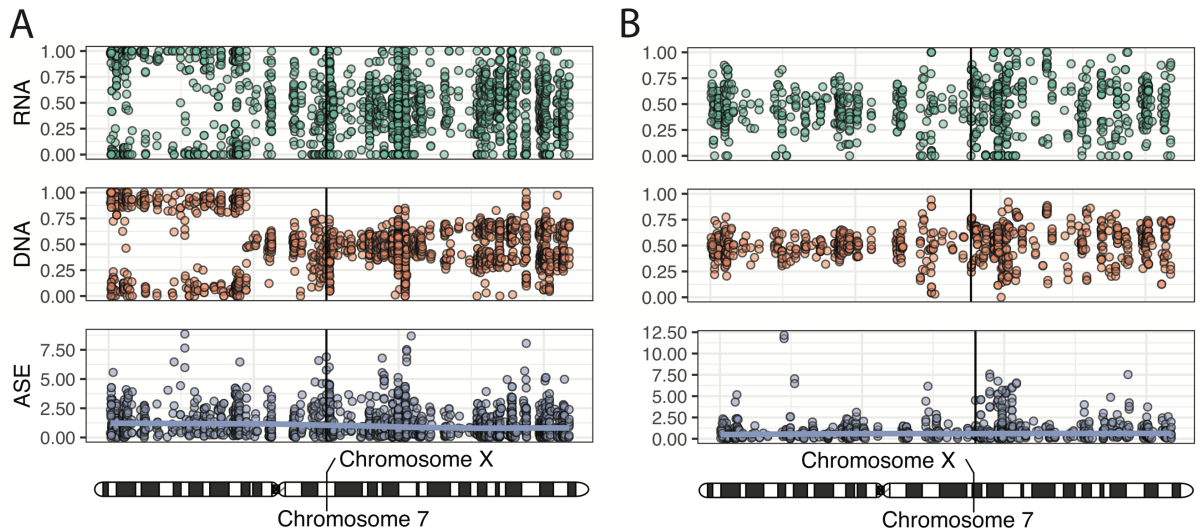


Figure 4: Some Translocations of the inactive X chromosome Do Not Suggest Autosomal Spread of XCI. The alternative allele fraction detected in RNA-seq and WGS as well as the corresponding log odds ratio for ASE were computed for all heterozygous germline SNPs on the autosomal partner chromosome of active X-chromosomal translocations. The average log odds ratio is indicated by the purple spline fit and the breakpoint of X-chromosomal translocations is highlighted in black. **(A)** LOH on chromosome 7 sufficiently explains the observed ASE after translocation of the inactive X chromosome in an OSC patient. **(B)** A copy number gain of chromosome 7 after translocation of the inactive X chromosome causes low levels of ASE in an HCC patient but does not suggest autosomal spread of XCI.

For the GBM and BLCA patient, long-ranging ASE in absence of additional structural variants could be observed for the autosomal chromosome arms in adjacency to the inactive X chromosome (Figure 5 and 6). A balanced 1q;X translocation in the GBM patient resulted in strong ASE across most of chromosome 1q and only declined close to the centromeric region. Notably, no other structural variation was detected on chromosome 1q for this patient (Figure 5A and Figure 6A). For the BLCA patient, a 21;X translocation but no other structural variation on chromosome 21 could be detected. Only a few protein-coding genes reside on chromosome 21, but a long-ranging ASE signal was observed (Figure 5B and 6B). Since translocations involving the inactive X chromosome were the only events on the relevant autosomal segments for both of these patients and ASE could be detected up to 50 MB from the translocation breakpoint, these cases potentially represent autosomal spread of X-silencing after somatic translocations.

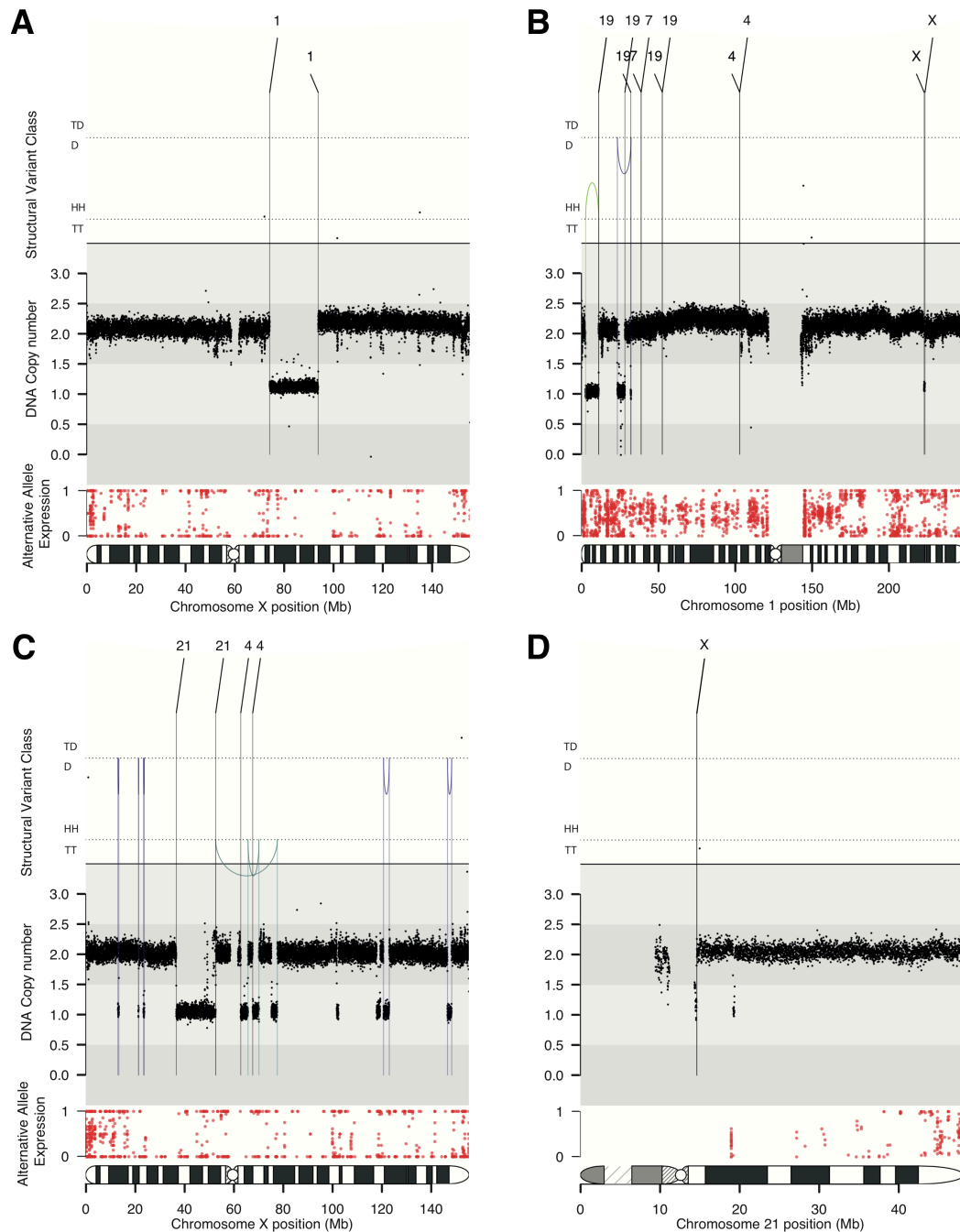


Figure 5: Translocations in a Glioblastoma and a Bladder Urothelial Cancer Patient Involved the Inactive X-Chromosomal Copy. TD: Tandem duplication. D: Deletion. HH: Head-to-head inversion. TT: Tail-to-tail inversion. **(A & B)** Glioblastoma patient. **(A)** Translocation to chromosome 1 coinciding with copy number loss but maintained ASE identified the involvement of the inactive X-chromosomal copy in a balanced 1q;X translocation. **(B)** Reciprocal translocation to chromosome X was the only structural variant detected on chromosome 1q. **(C & D)** Bladder urothelial cancer patient. **(C)** Copy number loss coinciding with the 21;X translocation and maintained ASE on chromosome X suggested translocation of the inactive X-chromosomal copy. **(D)** Translocation to chromosome X was the only structural event detected on chromosome 21.

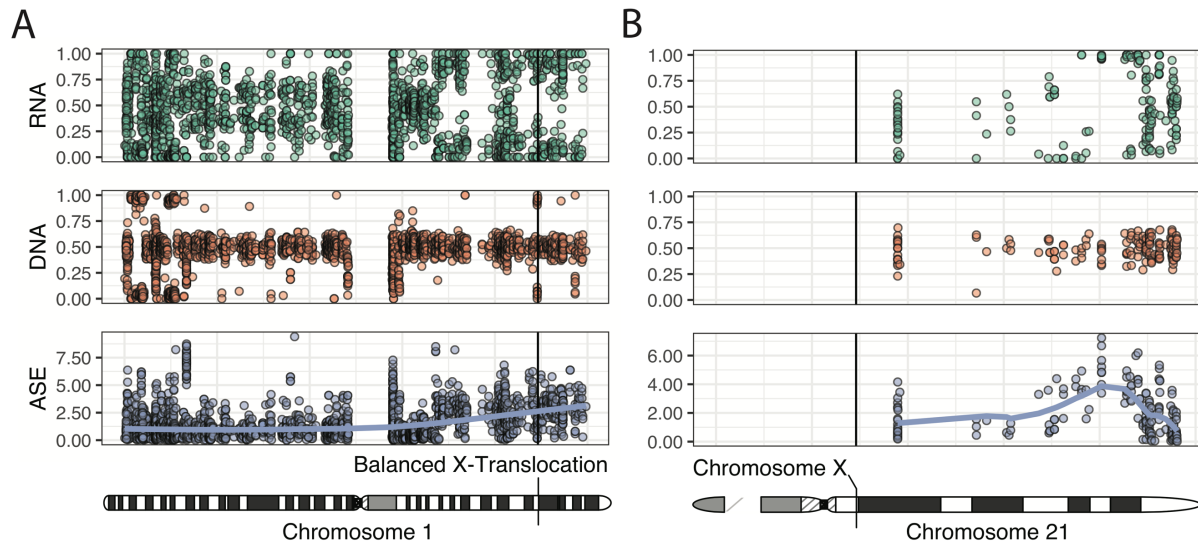


Figure 6: Autosomal Spread of XCI After Translocation of the Inactive X Chromosome is Suggested in a Glioblastoma and a Bladder Urothelial Cancer Patient. The alternative allele fraction detected in RNA-seq and WGS as well as the corresponding log odds ratio for ASE were computed for all heterozygous germline SNPs on the autosomal partner chromosome of inactive X-chromosomal translocations. The average log odds ratio is indicated by the purple spline fit and the breakpoint of X-chromosomal translocations is highlighted in black. **(A)** Substantially elevated ASE is detected on chromosome 1q after balanced translocation of the inactive X chromosome in a GBM patient. **(B)** Translocation of the inactive X chromosome causes chromosome-wide ASE on chromosome 21 in a BLCA patient after an unbalanced translocation of the inactive X chromosome.

Autosomal Segments Affected by Long-ranging ASE Adjacent to Inactive X-Chromosomal Translocations Display Reduced Gene Expression

Physiological inactivation of the X chromosome in women during development represents a dosage compensation mechanism¹⁷³. Consequently, chromosomal segments affected by autosomal spread of XCI were expected to be downregulated in addition to the observation of ASE. To corroborate the potential spread of XCI on chromosome 1q in a GBM patient and on chromosome 21 in a BLCA patient, the expression of these genomic regions was compared to the average expression of the corresponding cancer type.

For all genes that were expressed for the corresponding sample with observed ASE, the average FPKM values were computed from all other PCAWG donors with RNA-seq data of the appropriate cancer subtype. When the subtype-specific gene expression was compared for the relevant genomic segments, significantly lower

levels of expression was observed for both the GBM and the BLCA patient (Figure 7; $p = 1e-11$ for GBM and $p = 5e-4$ for BLCA, Wilcoxon-Mann-Whitney test). Since the same regions were affected by reduced expression in addition to ASE after somatic translocation of the inactive X, this further substantiated the evidence for autosomal spread of X-silencing in the GBM and BLCA patient.

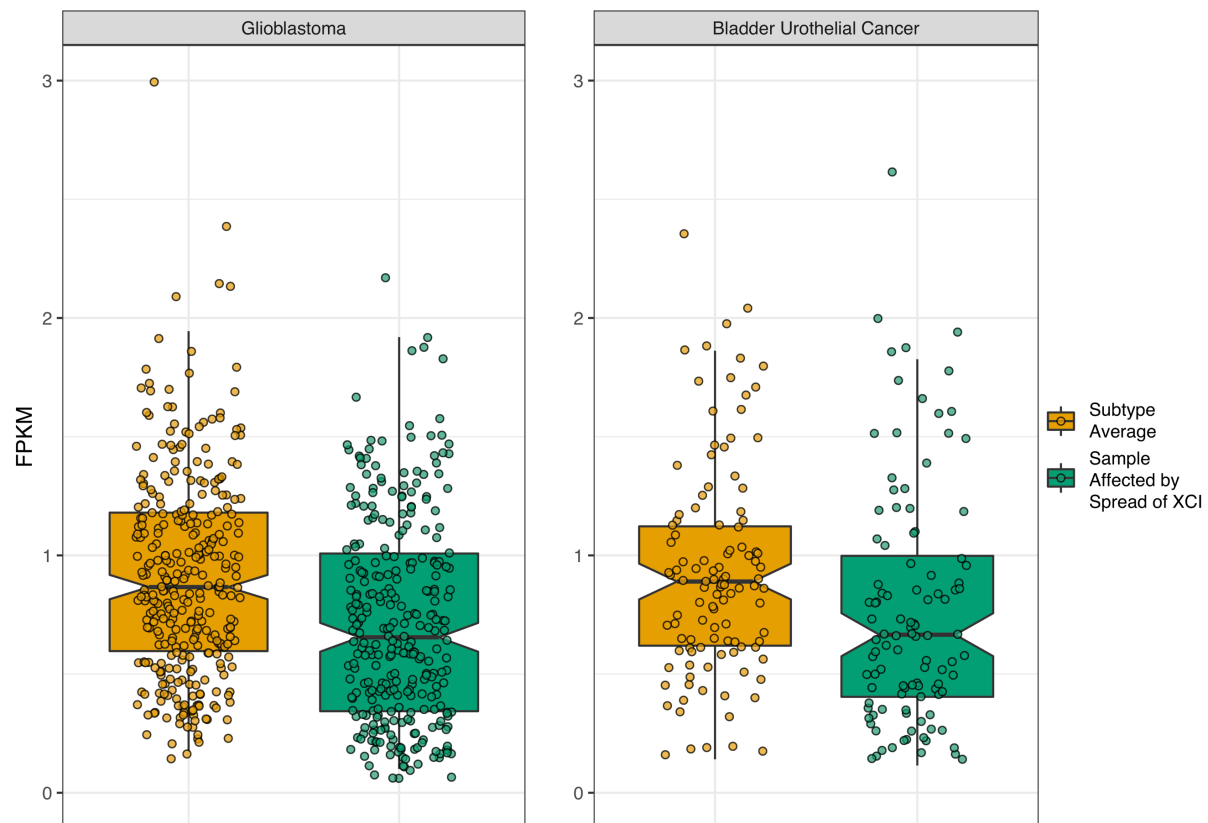


Figure 7: The Glioblastoma and Bladder Urothelial Cancer Patient Display Significantly Reduced Expression in the Autosomal Segment Affected by ASE After Translocation of the Inactive X Chromosome. Expression of protein coding genes on chromosome 1q for the GBM patient and chromosome 21 for the BLCA patient were compared to the average expression of the corresponding cancer subtype. For both patients, the chromosomal segments affected by ASE after translocation of the inactive X chromosome displayed significantly lower expression than the subtype-specific average ($p = 1e-11$ for GBM, $p = 5e-4$ for BLCA, Wilcoxon-Mann-Whitney test).

Discussion

Translocations are recognised as an important class of structural events in cancer that can impact genome stability in several ways^{138,140}. While disease-relevant translocations are typically linked to the generation of fusion genes, disruption of tumour suppressor genes or enhancer hijacking, the presented data highlights autosomal spread of XCI as additional mechanism for long-ranging dysregulation of gene expression. Autosomal spread of epigenetic silencing is established for germline disorders but only one case study could demonstrate this for a somatic translocation in a myeloproliferative neoplasm^{155-161,163}. Here, a large pan-cancer data set is systematically screened for X;autosome translocations and potential cases of autosomal spread of XCI.

From 620 considered female cancer patients, X;autosome translocation events could be identified in 158 samples. An initial comparison of ASE in proximity of all translocation-annotated breakpoints suggested an enrichment in adjacency to X;autosome translocations (Figure 1). Notably, this difference was detected regardless of the activity status of the X chromosome involved in the translocation. The assignment of the X-activity status was only possible in a subset of twelve manually curated patients (Figure 2). The absence of epigenetic information and the small subset of patients that could eventually be considered for a detailed analysis, also represent the main limitations of this study. However, two cases in which somatic translocation of the inactive X chromosome caused long-ranging ASE and downregulation of gene expression on the autosomal translocation partner could be identified by combining genome and expression data (Figure 4-6). The similarity to physiological X-silencing and the absence of any other genomic alteration strongly suggests that these cases are genuine examples of autosomal spread of XCI. The occurrence in two different genomic regions and tumour types within a large pan-cancer data set highlights this mutational mechanism as a rare but recurrent event in cancer.

Integration of Multiple Displacement Amplification into G&T-seq

Introduction

Multiple Displacement Amplification (MDA) is one of several WGA methods that are necessary to generate sufficient input material from the DNA of single cells for sequencing experiments ^{105,174-176}. G&T-seq is a modern multiomics technique that allows for the parallel analysis of the genome and the transcriptome from the same single cell ^{177,178}. An initial nucleic acid separation step within G&T-seq necessitates the DNA amplification on magnetic beads and although several MDA kits with slightly different chemistries exist, none of them was optimised for this particular experimental design. Therefore, this chapter demonstrates the suitability of MDA for G&T-seq and selects the optimal available chemistry for this approach.

A brief overview of single cell transcriptome and genome sequencing as well as relevant multiomics approaches is presented to provide a background for this rapidly evolving field and to highlight the importance of method and chemistry validation within new experimental designs.

Single Cell Transcriptome Sequencing

The transcriptome was the first entire molecular layer that single cell sequencing methods were presented for in 2009, only one year after corresponding bulk sequencing approaches were published ¹⁷⁹⁻¹⁸¹. While many different methods with complementary features exist nowadays, they all comprise the initial isolation and lysis of the cell or nucleus, followed by reverse transcription (RT) of the RNA and a subsequent amplification of the cDNA ¹⁸². Over time, novel methods have substantially reduced the levels of technical noise through optimisation of the involved enzymes, primer sequences or reaction conditions and also increased the throughput by introducing cellular or molecular barcoding techniques that allow for pooled amplification (reviewed in ¹⁸³). In general, the approaches can be divided into plate-based and droplet-based methods. Plate-based approaches typically rely on the physical separation of cells by micromanipulation, limiting dilution or fluorescence-

activated cell sorting (FACS). The wells contain barcoded oligos that are added through RT, ligation or PCR, which improve the throughput by allowing for pooled reactions of subsequent steps ^{182,183}. While plate-based formats are intrinsically limited in throughput, modern approaches using microwell-plates, droplet-containers or sequential combinatorial indexing can yield up to a few tens of thousands of cells per experiment ^{184,185}. Droplet-based methods use microfluidics to capture individual cells in nanolitre emulsions in which the initial barcoding reactions are performed. In contrast to the plate-based methods, there is no intrinsic throughput limitation for droplet-based methods but typical experiments aim to profile thousands to tens of thousands of cells per run ¹⁸³. Another important distinction of single cell RNA-seq (scRNA-seq) technologies that is also correlated to its throughput, is the difference between tag-based profiling and sequencing of full-length transcript structures ¹⁸². Sequencing of the complete transcript structure allows for the detection of alternative splicing, population-based genotyping as well as the estimation of RNA editing and allele-specific expression (ASE). However, despite a rapid decline of sequencing cost, full-length and high-depth analysis of the transcriptome from a large number of cells is still economically challenging and thus, several methods only aim to obtain short fragments from the 5'- or 3'-ends of the transcript. While losing some granularity, these tag-based methods enable the global characterisation of gene expression profiles of a large number of cells ¹⁸².

The availability of scRNA-seq revolutionised the way cell types are defined nowadays. Traditionally, cell types were defined by their location within tissues and their morphological features. The advent of monoclonal antibodies and FACS as well as Fluorescence *in situ* hybridisation (FISH) made it possible to detect particular proteins and surface antigens, genomic features such as rearrangements as well as specific RNA transcripts and added further resolution to cell type definition (reviewed in ¹⁸⁶). While these approaches already associated cell types with the presence of molecular markers, they are based on sporadic discoveries and not the result of systematic and comprehensive mapping of entire molecular layers. In contrast, scRNA-seq and genome-wide expression profiling allow for the unbiased classification of cells with similar molecular phenotypes and thus, are the current approach for cell type classification ¹⁸⁶.

Despite its great potential and widespread use, scRNA-seq comes with inherent limitations. For example, most techniques only achieve to capture 3 – 20% of all messenger RNA (mRNA) molecules per cell, leading to high technical variability between cells ¹⁸⁷. Additional technical variation such as batch effects and amplification biases together with intrinsic biological variation such as stochastic gene expression or cell cycle-based fluctuations, pose additional challenges to the robust analysis of scRNA-seq data ¹⁸². However, these limitations are increasingly recognised and addressed with active bioinformatic method development ¹⁸⁸⁻¹⁹⁰. While several other molecular layers can be probed with single cell sequencing nowadays and some remaining limitations of scRNA-seq, transcriptome analysis is still considered the most mature single cell technology ¹⁹¹.

Single Cell Genome Sequencing

Single cell genome sequencing is a more recent technology compared to single cell transcriptomics. While it was already shown that amplifications of whole human genomes were feasible in 2002 and NGS technology became available in 2007, the first single cell whole-genome sequences from human cells were published in 2011 ^{50,175,192}. In this work, 200 single nuclei from breast tumours and their liver metastases were isolated and despite only covering about 6% of the genome per cell, extensive subclonal copy number evolution could be demonstrated¹⁹². In the following years, single cell genomics contributed substantially to our understanding of genomic heterogeneity in cancerous and normal human tissues ^{67,107,176,193}.

Diverse classes of mutations such as retrotransposition, CNVs or SNVs have been analysed in various tissues with single cell genomics. Similar to single cell transcriptomics, many different single cell genome sequencing methods exist ¹⁷⁶. In general, they all include the common steps of isolation and lysis of cells or nuclei, followed by WGA before sequencing library preparation. WGA is necessary as single cells only contain around six picograms of DNA, which is orders of magnitude below the required input for sequencing experiments ¹⁷⁶. Several different WGA methods have been developed that all have unique benefits but also an associated spectrum of artefacts and technical limitations. Thus, the choice of the WGA method is crucial and limits the eventually generated sequencing data to certain downstream analyses ^{105,106,176}.

Extensive efforts have been made to compare various WGA methods and chemistries^{105,106,194-200}. Usually, these comparisons are aiming to assess the suitability of a range of commercially available WGA methods for particular or several downstream analyses^{105,106,197-199}. Moreover, they are trying to assess the impact of the quality or different sample states such as fresh or frozen tissue on different WGA techniques or focus on a particular organism or tissue^{194,195,200}. WGA methods are often classified into three groups and while general trends for each class emerge from these comparisons, it is also evident that the obtained results differ based on the precise application and experimental conditions. Notably, differences can be detected between kits that rely on the same basic class of WGA but are from different commercial suppliers^{105,197}. Therefore, it is crucial to validate the WGA of choice for new experimental conditions. To provide a general overview about different WGA classes and their advantages and limitations, a brief summary is given below.

First approaches completely relied on PCR amplification that target common or adaptor sequences or used degenerate-oligonucleotide primed PCR (DOP-PCR; Figure 8A)^{105,201}. While the first single cell genomics study successfully applied this WGA method on tetraploid nuclei to study large-scale copy number variation, DOP-PCR generally suffers from poor genome coverage, high allelic dropout (ADO) and an elevated point mutation error rate compared to more recent approaches^{105,106,176,202}.

To mitigate the low breadth of genome coverage when solely relying on PCR amplification, a more diverse class of methods uses an initial strand displacement amplification step to introduce an adaptor onto the copied strands. This limited isothermal amplification is commonly followed by several rounds of semi-linear amplification of the initial product to improve uniformity across the genome before a final PCR-based amplification^{202,203}. Since they combine isothermal displacement with PCR-based amplification, they are referred to as hybrid methods (Figure 8B). These hybrid methods have been shown to provide the highest coverage uniformity of currently available WGA methods with significantly increased breadth of coverage compared to DOP-PCR^{105,106,176,197}.

The third class of WGA comprises multiple displacement amplification using isothermal reactions with the high-fidelity Φ 29 DNA polymerase and random

primers^{174,175,204}. Through a continuous process of strand displacement and random priming, MDA generates highly branched structures in an exponential amplification process (Figure 8C). This exponential amplification can lead to pronounced non-uniformity of the amplified DNA resulting in high levels of ADO and artificial biases for copy number detection^{105,106,194,204}. While less suitable for the detection of CNVs, MDA is preferred for genome-wide single cell SNV analyses due to its superior breadth of coverage and superior proof-reading capacity of the $\Phi 29$ DNA polymerase^{105,106,176,197,200}.

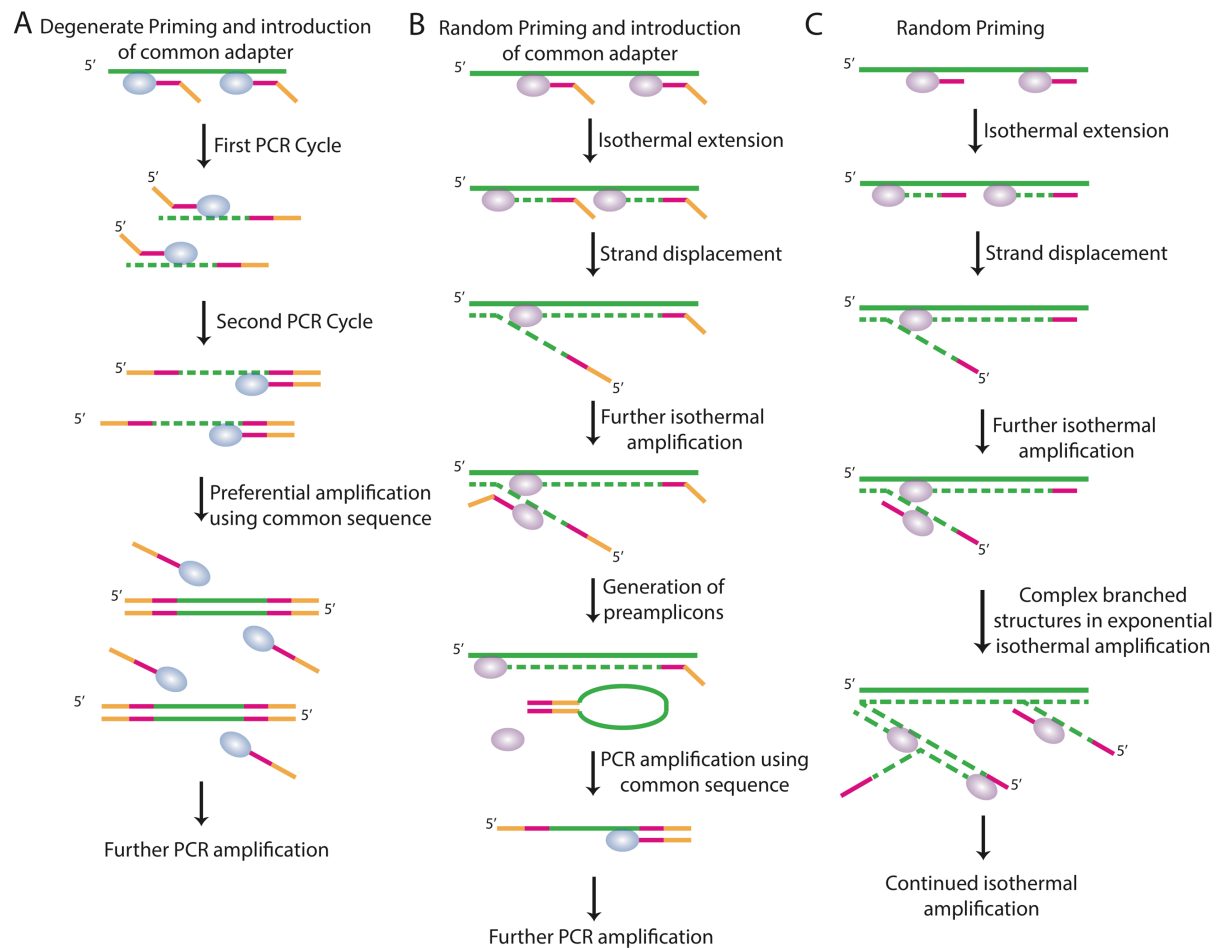


Figure 8: Schematic Overview of WGA Methods. (A) Pure PCR Methods. PCR methods prime on naturally occurring or introduced common sequences and achieve WGA through continuous PCR amplification. **(B) Hybrid Methods.** Hybrid methods use initial (semi-)linear strand displacement amplification before a PCR-based amplification. **(C) MDA.** MDA solely relies on the isothermal strand displacement amplification.

While MDA represents the preferred WGA method for the detection of base substitutions, it is important to note that heat-dependent cytosine deamination occurs

during common lysis protocol and the subsequent enzymatic amplification still introduces an average of 1000 erroneously paired bases for every complete amplification of the human genome ^{105,106,204}. This remaining burden of technical artefacts can be aggravated by the non-uniform amplification of MDA, which can make it difficult to distinguish between early introduced artefacts and true biological variants ⁵¹. These limitations were targeted by several technical improvements of the method such as alkaline cell lysis on ice or the potential for molecular barcoding ^{104,205}. Furthermore, several bioinformatic tools were developed ^{51,104,206}.

The first SNV studies in single cells used traditional variant callers that were designed for bulk sequencing that do not specifically address the technical limitations of single cell data ^{103,202,205,207}. Since then, several single cell-specific SNV callers have been developed. One class of tools utilises shared information across several cells while either modelling single cell-specific error profiles or trying to reconstruct consistent phylogenies ²⁰⁸⁻²¹¹. However, these approaches are blind to private mutations and less applicable in highly heterogeneous or post-mitotic tissues. Other algorithms achieve improved accuracy through the integration of heterozygous germline SNPs. SCCaller derives a local estimate of the amplification non-uniformity and aims to threshold between artefacts and true variants based on this estimate and the observed variant allele fraction of potential SNVs ¹⁰⁴. LiRA uses read-backed phasing and requires absolute concordance between the observed alleles of germline and potential somatic variants ²⁰⁶. While these approaches can detect private mutations, their thresholding approaches still fail to remove artefacts that were introduced in early rounds of MDA. Consequently, SNV calling in single cells is still a field under active development and in need for new approaches to distinguish between MDA-based artefacts and genuine biological mutations.

Single Cell Multiomics

Single cell sequencing methods have not only been developed to characterise the transcriptome or genome but also several epigenome layers such as the methylome and chromatin accessibility profile. Beyond that, several current technologies allow for the parallel measurement of two or even three molecular layers from the same single cell. These single cell multiomics methods allow for the direct correlation between the

measured layers without being obscured by cellular heterogeneity and have the potential to unravel their complex interaction networks (reviewed in ^{212,213}).

Currently, there are three main methods that enable transcriptome and whole-genome sequencing of the same single cell ^{177,214,215}. All of these methods have primarily been used to investigate the impact of aneuploidies, CNVs and inter-chromosomal fusions on gene expression and could demonstrate a positive correlation between these measurements ^{178,214,215}. While the demonstrated effects are similar, their approach to derive the whole genome and transcriptome differs.

In DR-seq, cells are manually isolated, lysed and oligo-dT primed reverse transcription introduces cellular barcodes, an adapter sequence and a particular a T7 promoter on the first-strand copy DNA (cDNA) in the presence of genomic DNA (gDNA). Subsequently, the cDNA as well as the gDNA are subjected to a quasilinear amplification that introduces a second adapter. The amplified product is split before the cDNA is further amplified using *in-vitro* transcription to ensure the absence gDNA contamination. In parallel, the gDNA is amplified using a PCR priming on the adapter introduced during the quasilinear amplification. However, cDNA molecules that are present in the gDNA amplification reaction carry one or two adapters and thus, can be co-amplified with the targeted gDNA ²¹⁴ (Figure 9A). While the amplification process is straightforward for DR-seq and loss of original gDNA and RNA molecules is minimised due to the initial co-amplification, the choice of the WGA method is not flexible, WGA-related artefacts are present in the cDNA, which only covers the 3'-ends due to *in-vitro* transcription and the genome sequencing data features cDNA contamination.

The other methods both rely on physical separation of the DNA and RNA before their amplification. For SIDR-seq, isolated cells are incubated with antibody-conjugated magnetic beads that bind to the cellular membrane. The subsequent gentle lysis maintains the nuclear integrity. Thus, the nucleus can be precipitated in a magnetic field while the supernatant containing the cytoplasmic RNA can be removed. In theory, any whole transcriptome amplification (WTA) and WGA method available for single cells can be used afterwards due to their physical separation (Figure 9B) ²¹⁵. While physical separation allows for a more flexible amplification approach, SIDR-seq loses

the nuclear RNA fraction and is dependent on tissues, where whole cell isolation is feasible.

G&T-seq is the first method that utilised the idea of physical separation of the molecular layers before sequencing to allow for a flexible amplification approach. After cellular lysis, the mRNA is precipitated using oligo-dT-coated paramagnetic beads and the DNA-containing supernatant is transferred to an empty multi-well plate using liquid handling robotics. Following a bead-based precipitation of the DNA in the transferred supernatant, any WTA and WGA protocol can be applied in principle (Figure 9C)^{177,178}. In contrast to DR-seq and SIDR-seq, G&T-seq is independent of a particular cellular isolation method and works equally well on whole cells or isolated nuclei. This is particularly useful for tissues or cell types, where isolation of whole cells is not feasible.

While any WGA method is compatible with G&T-seq in theory, only hybrid class approaches have been tested thoroughly as previous efforts were focussed on CNV analyses¹⁷⁸. Furthermore, the precipitation and transfer of DNA before WGA could consistently lead to the loss of DNA molecules, preventing the amplification of homologous alleles and limiting the breadth of coverage. Any additional processing step also bears the potential to introduce technical artefacts, which could affect the observed mutational burden and spectrum. To extend the applicability of G&T-seq and to offer a multiomics approach to SNV calling in single cells, a comprehensive MDA chemistry comparison within the G&T-seq workflow was performed and is presented in this chapter.

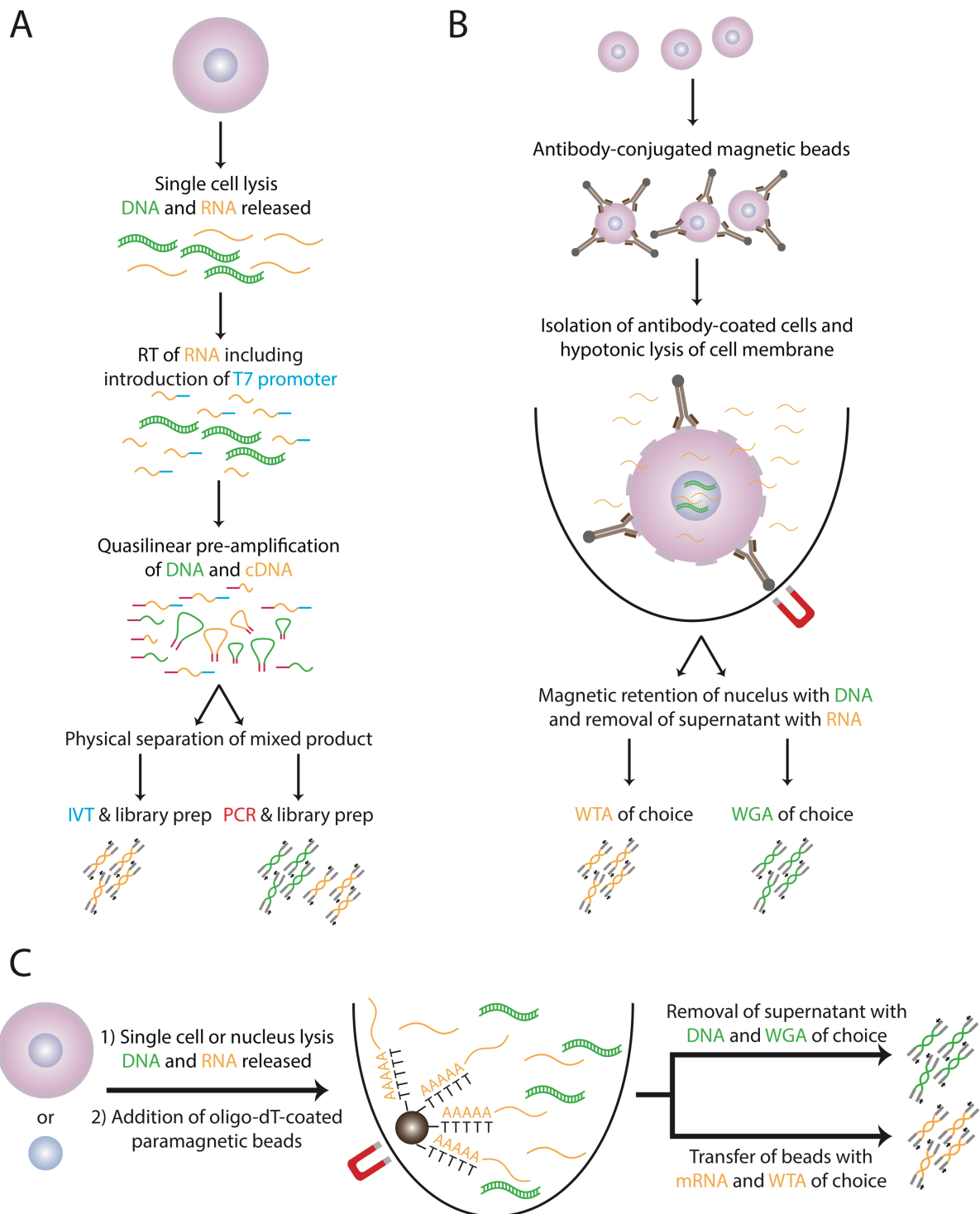


Figure 9: Parallel DNA-and-RNA Single Cell Sequencing Techniques. (A) **DR-seq.** DR-seq performs RT and a semi-linear amplification in presence of both nuclear acids before physically splitting the product. Therefore, the library submitted for WGS can contain cDNA amplicons. (B) **SIDR-seq.** SIDR-seq separates RNA from DNA through magnetic retention of the nucleus containing the genomic DNA. (C) **G&T-seq.** Both single cells or single nuclei can be used as input material to G&T-seq. Physical separation of DNA and RNA before amplification is achieved through magnetic retention of polyA-tailed mRNA.

Methods

Cell Culture and Single Cell Isolation Using FACS

Cell culturing was performed by Iraad Bronner within the Core Facility Pipelines at the Wellcome Sanger Institute (Sanger Pipelines) and single cell isolation was done in cooperation with Iraad Bronner and the Cytometry Core Facilities at the Wellcome Sanger Institute. Low-passage aliquots from the cell line HCC38-BL were thawed from liquid nitrogen and maintained in standard cell culture conditions for a few days to allow one passage and subsequent outgrowth. Cells were suspended using a brief Trypsin digestion and cell clumps were removed using a cell strainer. Hoechst stain was added to the cell suspension and 50-cell controls as well as single cells were sorted into 96-well plates containing 1-2.5 μ L of the lysis buffer or PBS as appropriate for the subsequent MDA method using the BD Influx Cell Sorter from BD Biosciences. Doublet discrimination was performed using a Forward-Scatter height versus Forward-Scatter area approach. The cell cycle states G1 and G2 were discriminated by their DNA content as indicated by the Hoechst staining. A very narrow gate around the two peaks observed at DNA contents at 100% and 200% (arbitrary units) was chosen to specifically select for cells in G1 or G0 and G2 or M phase while avoiding cells that are undergoing DNA replication during S-phase. The 96-well plates were briefly spun down and stored at -80 °C until further processing.

MDA Amplification and G&T-seq

Three commercially available MDA chemistries were selected for the comparison between the G&T-seq and the respective original single cell genome sequencing version. The chemistries selected were the Single Cell GenomiPhi DNA Amplification kit from GE Healthcare Life Sciences (GenomiPhi), the Repli-g Single Cell Kit from Qiagen (RepliG) and the Trueprime Whole Genome Amplification Kit from Expedeon (Trueprime). All processing steps within G&T-seq and MDA amplifications were performed by Iraad Bronner. The commercial MDA kits were delivered with their individual lysis buffer, while the Buffer RLT Plus from Qiagen was used as lysis buffer within G&T-seq. The 96-well plates containing cells sorted into Buffer RLT Plus were subjected to the separation of mRNA and gDNA and the gDNA was transferred and precipitated as described in the original publications^{177,178}. Briefly, the polyadenylated mRNA was hybridized with paramagnetic beads carrying modified oligo-dT primers.

During magnetic retention of the bead-bound mRNA, the original supernatant containing the gDNA and four washing step supernatants were pooled in a new 96-well plate. Subsequently, gDNA was precipitated within this new plate using AMPure Beads and washed with ethanol. After the ethanol dried off, the beads and gDNA were resuspended in the appropriate volumes of PBS and MDA reaction buffer as per the manufacturer's instructions for the three different MDA chemistry kits. Similarly, the respective amounts of lysis or reaction buffers were added to the wells subjected to the original MDA process. For each of the considered commercial MDA kits, the MDA reactions were performed in parallel for the G&T-seq and the corresponding original version. Besides the mainly non-disclosed differences in buffer composition, the three MDA kits also differed in the suggested duration of MDA. MDA was performed at 30 °C for 2 h for GenomiPhi, for 3 h for Trueprime and for 8 h for RepliG followed by a brief enzyme inactivation at 65 °C for all three kits. Subsequently, the MDA products were purified using a standard Ampure SPRI bead clean-up and quantified using the spectroscopic PicoGreen method ²¹⁶.

Genotyping-by-Sequencing for MDA Product Quality Control

The Genotyping-by-Sequencing (GbS) assay as developed within the Sanger Pipelines employs a bespoke miniaturised high-throughput amplicon sequencing protocol. All GbS-related lab work for the present samples was performed in cooperation or solely by Scott Goodwin and Naomi Park within the Sanger Pipelines. GbS is a high-multiplex two step PCR, with the first PCR stage employing bipartite primers to capture 126 SNP sites and their surrounding regions (190-250 bp typical size) and incorporate Illumina sequencing motifs, and the second PCR stage also employing bipartite primers which target the previously-introduced sequencing motifs and introduce dual-index barcodes and Illumina flowcell adaptor sequences. The assay is performed in a multi-well plate setup at a final volume of 5 µL and requires a minimum input of only 1 ng of DNA.

The first PCR stage deviates significantly from standard PCR practice in order to ensure that targets are captured with approximately equal efficiency and with minimal off-target amplification. The primers were designed using the MPprimer software tool and feature 2'-OMe modifications at the penultimate 3' base to inhibit primer-dimer formation ²¹⁷. They are included in the PCR at average concentrations of 300pM, with

the exact concentrations of individual primer pairs being slightly variable to compensate for amplification efficiency differences. The first PCR stage is composed of 5 elongated cycles of 40 minutes duration to permit efficient target capture in spite of the low primer concentration. At the end of this stage, the reactions are cooled to 4°C to temporarily inhibit polymerase amplification and the second stage barcoding primers are introduced. Due to the low reaction volumes it is not possible to introduce the second stage primers into the PCR in solution form, therefore instead the entire reaction volume is transferred into another plate which contains pre-aliquoted dried barcoding primers in each well, and the reactions are mixed thoroughly to rehydrate and disperse the barcoding primers.

The second PCR stage follows a conventional format and duration with 200nM primer concentration and typically 31 cycles for a total duration of approximately 1 hour. The barcoding primers amplify only those products generated in the first PCR by targeting the sequencing motifs introduced by the first stage primers. The PCR conditions have been chosen to achieve two aims; firstly that the carryover first stage primers no longer significantly participate in the PCR (there is approximately a 1000x excess of barcoding primer compared to each individual first stage primer pair, and annealing times are insufficient to allow efficient amplification by the first stage primers), and secondly that the amplification achieves plateau by effective exhaustion of the available barcoding primers over a wide range of input, therefore total yields per PCR are normalised to approximately the same level irrespective of the initial DNA concentration.

Therefore, amplicons can be pooled in an equivolume manner without any further manipulation as normalisation for both amplicon yield and sample yield is achieved within the PCR. Pooling is done by inverting the open PCR plates over a specially-designed reservoir and gentle centrifugation. Aliquots of this pool are purified using a standard Ampure SPRI bead clean-up on a liquid handling robot, followed by quantification by qPCR and sequencing on one Illumina MiSeq lane for 384 cells.

The sequencing data was returned as cram files aligned to GRCh37. Computational evaluation of the GbS data was done by myself. A pileup across the 126 target regions was performed for the MDA products from this comparison experiment as well as for

HCC38-BL bulk sequencing data that was already available within the Wellcome Sanger Institute using bcftools mpileup and SNPs were called using bcftools call ²¹⁸. Any calls with a total depth of less than six reads or phred-scaled quality score of less than 30 were flagged. To assess the quality of individual MDA products, their genotypes were compared to the reference genotypes that were called in the bulk sequencing data. For homozygous SNPs in the bulk sample, the MDA-derived genotypes were classified as concordant if they matched the SNP called in the bulk sample, as dropout if no sufficient coverage was obtained or as false positive if a non-matching genotype was called. Additionally, ADO was classified for heterozygous SNPs in the bulk sample if the MDA-derived product was called to be homozygous for one of the respective alleles in the bulk sample. The GbS results were used to select high-quality MDA products for WGS.

Sequencing Library Preparation and WGS Sequencing

Amplification-free paired-end sequencing libraries for the selected MDA products were created in the Sanger Pipelines using standard Illumina protocols ²¹⁹. Samples were multiplexed in equimolar amounts and submitted for 150 bp paired-end sequencing on the Illumina HiSeq X platform within the Sanger Pipelines.

WGS Quality Control and Genome Alignment

The WGS data was returned as de-multiplexed cram files per HiSeq X lane. The cram files were converted into FASTQ format using biobambam ²²⁰. FASTQ files from all HiSeq X lanes per MDA product were concatenated and Illumina sequencing adapter, unknown nucleotides and bases with phred-scaled quality scores less than 10 were removed from the read ends using TrimGalore and Cutadapt ²²¹. Paired-end alignments to GRCh37 were performed using bwa mem with default parameters ²²². Genome alignments were sorted and indexed using samtools and duplicates were removed using Picard v1.6 ^{223,224}.

Computation of Depth, Breadth and Uniformity of Genome Coverage

A genome-wide computation of sequencing depth per base was performed using samtools depth ²²³. The average depth as well as the cumulative fraction of the genome covered at various coverage thresholds was computed in R using this data. Lorenz curves to illustrate the uniformity of coverage were obtained using the R

package `ineq` and Gini coefficients were computed using the `auc` function from the R package `MESS` ^{225,226}.

Genotyping of MDA-derived Whole-Genome Sequences

A list of 2,011,109 frequently polymorphic sites was used to obtain reference genotypes from the bulk sequencing data using the `bcftools mpileup` and `call` commands. A minimum coverage of six reads and a phred-scaled quality score of at least 30 was required to keep only high-quality reference genotypes. Genotypes of MDA-derived whole-genome sequences were obtained using similar commands but restricted to the 1,878,813 sites with known reference genotype and the minimum coverage was reduced to five reads and no quality-based filtering was performed to account for the lower average depth of MDA-derived whole-genome sequences.

Genome-wide SNV Calling

Genome-wide SNV calling was done using the single cell-specific caller `SCCaller-v1.2`. `SCCaller` derives a local estimate of the MDA amplification bias using known heterozygous SNPs from a corresponding bulk sample ¹⁰⁴. The approximately 516,000 SNPs that were called as heterozygous in the bulk data during the previous genotyping assessment were used for this. The amplification bias estimates can be refined if the amplicon size distribution of MDA products is known, as this is used to define the width and shape of the kernel smoother within `SCCaller`. Estimates of the size distributions were taken from the single cell WGS quality package `PaSD-qc` ²²⁷. To obtain these size distributions, the alignments were down-sampled to an average coverage of 1x using `samtools` and compared to the categorical spectra at the same coverage within the `PaSD-qc` pipeline ^{223,227}.

Given the locally observed MDA amplification bias as well as the total depth and variant allele fraction, `SCCaller` computes a likelihood parameter that aims to distinguish between technical artefacts and true variants ¹⁰⁴. All sites with a likelihood for being an artefact of greater than 0.01 as well as all variants present in dbSNP build 138 or the 1000 Genome Phase 1 panel were flagged and not considered in the final burden and mutational spectrum analysis.

Mutational spectrum analysis

Every mutation is the result of a corresponding mutational process as explained in the general introduction for this thesis. As mutational processes leave unique footprints, the observation of mutational spectra can give insight into biological processes as well as processing steps that cause technical artefacts. Traditionally, SNV calling results are defined as the base change on the Watson strand. As it is impossible to distinguish on which strand the original base change occurred after Illumina sequencing without specific barcoding approaches, all SNV calls were classified given the pyrimidine base change of the Watson-Crick base pair for the mutational spectrum analysis. Consequently, both an originally annotated G>A and an originally annotated C>T mutation, will be classified as C>T mutations for this analysis. This results in a total of six classes of base substitutions, namely C>A, C>G, C>T, T>A, T>C, T>G. This spectrum can further be refined by incorporating information about the 5'- and 3'-context of the mutated base. Here, a single base immediately prior and after the called SNV is considered, creating sixteen different permutations for a total of 96 substitution classes. This approach follows the commonly used pattern for mutational signatures of single base substitutions^{18,29,36}. The filtered genome-wide calls derived from SCCaller were classified accordingly and the resulting mutational spectrum plotted to illustrate any potential differences between the G&T-seq and the original RepliG version. No formal mutational signature extraction was performed.

Statistical Analysis

All statistical analysis was performed in R version 3.5.0 using core distribution functions unless otherwise indicated¹⁷².

Results

Experimental Setup

The main aim of this project was to extend the applicability of G&T-seq to SNV calling in single cells. While the final method should consist of a multiomics RNA-and-DNA calling approach, it was essential to provide evidence that SNV calls solely based on G&T-seq's genomics layer were comparable to conventional single-omics approaches. Since MDA-based methods are the current gold-standard for SNV calling in single cells, the WGA comparison was restricted to commercially available MDA chemistry kits ^{105,106,176}. The chemistries selected were GenomiPhi, RepliG and Trueprime as explained in the methods section above. Both GenomiPhi and RepliG use the traditional Φ 29 DNA polymerase with random hexamer priming and mainly differ in their undisclosed buffer composition. Trueprime uses the same DNA polymerase but contains a DNA primase instead of random oligonucleotide primers in order to improve evenness of amplification ²²⁸.

Within the scope of selecting the most suitable MDA chemistry for G&T-seq, there were several additional requirements. As PCR-based methods are more error-prone than MDA, the product yield had to be sufficient for amplification-free sequencing library preparation ^{176,219}. To perform these reactions within the Sanger Pipelines, a minimum of 500 ng was necessary. Since MDA products can be of variable quality and SNV analyses require moderately high sequencing depth, only high-quality amplifications should be subjected to whole-genome sequencing (WGS) for financial reasons. A previous publication demonstrated that a qPCR panel targeting a single locus per autosome could be used as quality control criterium for MDA based on the detection of all twenty-two qPCR products ²²⁹. To extend on this idea, a highly multiplexed PCR assay targeting several loci across all human chromosomes was going to be tested for its ability to predict WGS data quality. This assay can be followed up by shallow sequencing to check for the faithful amplification of a known reference genotype at a low cost for hundreds of cells at a time to select a subset of high-quality MDA products. Such an assay was in concurrent development in the Sanger Pipelines when the MDA chemistry comparison within G&T-seq was initiated and thus, was included in this workflow. The so-called GbS assay is described in the methods section in more detail. Briefly, it allows to assess the genotypes at 126 polymorphic loci sites

and is performed in a multi-well format on liquid handling robots which allows for high-scale throughput if desired. Finally, the impact of the cell cycle status was considered for any comparison as a previous publication indicated that a more even amplification is possible for cells in G2/M compared to G1/0 due to the inherently higher amount of genetic starting material ²²⁹.

The cell line HCC38-BL was selected for all amplification experiments as it had been used for the original G&T-seq publications ^{177,178}. Single cells and 50-cell controls were flow-sorted into two 96-well plates per MDA chemistry kit and classified for their G1/0 (only referred to as G1 hereafter) or late M and G2 (only referred to as G2 hereafter) cell cycle state using DNA content thresholds as indicated by Hoechst staining. Half of the plate was sorted into the lysis buffer from the MDA kit, while the other half was sorted into the G&T-seq appropriate lysis buffer and the bead-based separation of polyadenylated mRNA and gDNA for G&T-seq was performed only on this half of every plate. The DNA-fraction was transferred into a new multi-well plate and MDA was performed together with the original half of every plate according to the corresponding manufacturer's instructions. Flow-sorting, G&T-seq separation and MDA amplifications were done solely by or in cooperation with Iraad Bronner in the Sanger Pipelines and the Cytometry Core Facilities at the Wellcome Sanger Institute as indicated in the methods section. The general plate layout is displayed in Figure 10.

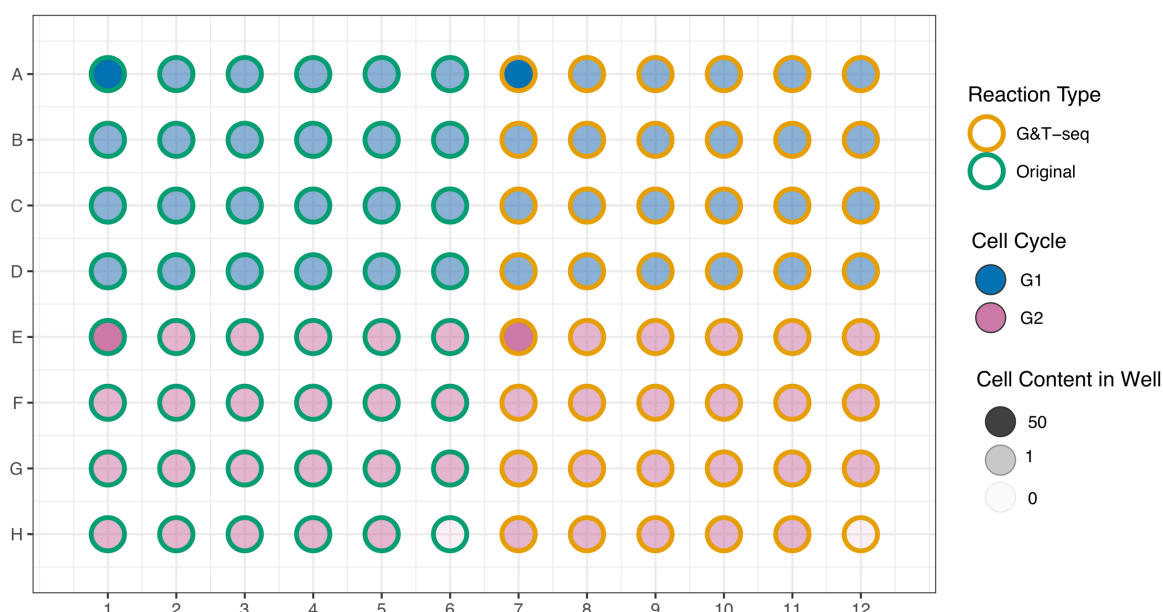


Figure 10: Plate Layout for MDA Chemistry Comparison. To minimise batch effects, half of the 96-well plate was FACS-sorted into buffer appropriate for either GenomiPhi, RepliG or Trueprime, while the other half of the plate was FACS-sorted into the G&T-seq appropriate lysis buffer. Additional G&T-seq processing steps were only performed on the appropriate half of the plate and the final MDA reaction was performed in parallel per originally sorted plate according to the respective manufacturer instructions. To investigate the difference of the cell cycle stage on MDA performance, half of the cells were selected to be in the G1 stage, while the other half was selected from the G2 stage based on the DNA content as indicated by Hoechst staining during FACS-sorting. For each of the four possible combinations, one 50-cell control was created and one well per reaction type was left empty to serve as negative control. Single cells were sorted into the remaining wells.

GenomiPhi Yields Insufficient Amounts of DNA

Subsequent to the MDA amplification according to the manufacturer's instruction, the product yield was evaluated using the PicoGreen spectroscopic method²¹⁶. This work was done in cooperation with Iraad Bronner from the Sanger Pipelines. The product yield was consistent within every condition and nearly identical between the G1 and G2 cell cycle state but showed great differences between MDA chemistry kits and the original or G&T-seq version (Figure 11).

For the original GenomiPhi version, an average product yield of 1.9 μg (SD = 1 μg) was obtained while the corresponding amplifications for G&T-seq only averaged 0.2 μg (SD = 0.2 μg). As the minimum yield for amplification-free sequencing library

creation was 500 ng within the Sanger pipelines, GenomiPhi did not reach this threshold within G&T-seq and was excluded from further comparisons (Figure 11).

For RepliG, the yield obtained from the G&T-seq version was comparable to the original one. They averaged 17.9 μg (SD = 3.7 μg) and 17.8 μg (SD = 5.6 μg), respectively and thus, significantly exceeded the minimum requirement of 500 ng (Figure 11). While the overall yield was lower, the G&T-seq and the original version were also comparable for Trueprime. The yield of 1.4 μg (SD = 0.6 μg) for the G&T-seq version as well as an average of 1.4 μg (SD = 1.1 μg) for the original version were also both suitable for the amplification-free sequencing library preparation (Figure 11).

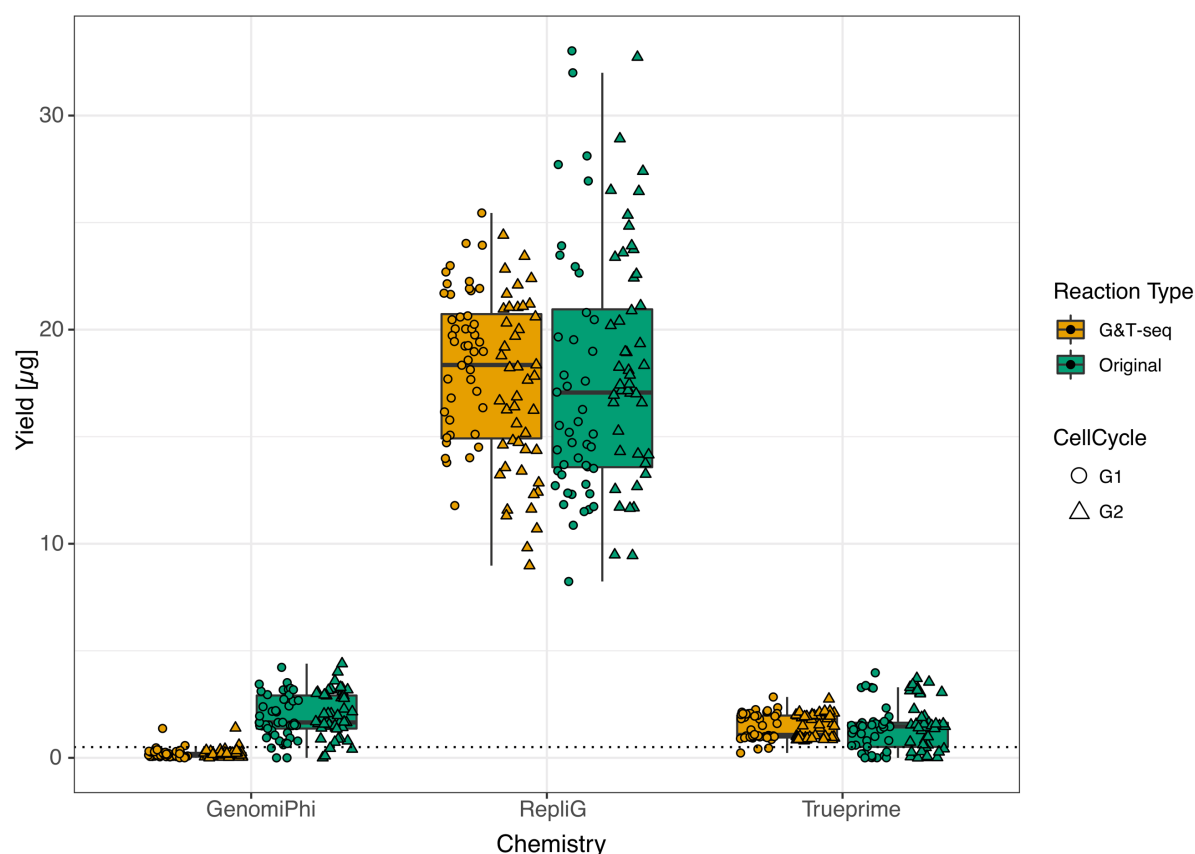


Figure 11: MDA Product Yield. The MDA product yield was determined using the spectroscopic PicoGreen method. A minimum of 500 ng was necessary for amplification-free sequencing library preparation within the Sanger Pipelines and is indicated with the dashed line. In general, the GenomiPhi reactions within G&T-seq fall short of this requirement while the original GenomiPhi as well as both the G&T-seq and the original versions for RepliG and Trueprime displayed sufficient yield. No difference could be observed between cells sorted for G1 or G2 cell cycle stage within the same reaction conditions.

Pre-sequencing Quality Control Suggests Poor Amplification for Trueprime

All RepliG and Trueprime amplification products were subjected to GbS and MDA-derived genotypes were called as described in the methods section. Reference genotypes were obtained from corresponding HCC38-BL bulk sequencing data that was available within the Wellcome Sanger Institute. The MDA-derived genotypes were compared to the reference genotypes and classified as concordant, dropout, false positive or ADO as explained in the methods section. In general, most cells displayed no false positive calls across the 126 tested SNP loci and only a minor number of ADO calls (Mean = 10, SD = 10). As cells with fewer ADO calls usually displayed a much greater amount of complete dropout, this did not indicate a more faithful but rather an overall less successful MDA. Therefore, dropout, false positive and ADO calls were pooled and directly contrasted with the number of concordant calls (Figure 12).

For RepliG, an average of 41% (SD = 17%) of the tested 126 SNP loci was called as concordant to the reference for the G&T-seq version and 30% (SD = 29%) for the original version (Figure 12). The lower average and greater deviation of concordant calls for the original version was explained by 39 of the 95 tested cells displaying a dropout rate of greater than 90%, indicating a complete failure of the MDA reaction for these cells. The cells amplified in the G&T-seq version displayed a much more uniform distribution around their average with only a single completely failed cell. As G&T-seq uses a harsher lysis buffer than any of the original versions of the MDA chemistry kits, the higher overall success rate for the G&T-seq version can potentially be attributed to a greater cell lysis efficiency. For successfully amplified cells, the distribution of GbS quality values was comparable between the original and the G&T-seq version and between the G1 and G2 cell cycle state.

MDA success as indicated by GbS was significantly worse for Trueprime products compared to RepliG (Figure 12). Only 5% and 2% of probed genotypes were called as concordant for the Trueprime G&T-seq and the corresponding original version, respectively. Even the most successful amplifications only reached 33% and 41% for the G&T-seq and the original version, respectively.

Based on the GbS assay, MDA products were selected for WGS. Since Trueprime amplifications were suggested to be of inferior quality compared to RepliG, only RepliG

products were chosen. A total of 20 MDA products were selected, evenly split between the original and the G&T-seq version as well as their cell cycle state. For all four conditions, the 50-cell control was chosen and the remaining four single cells were chosen across the range of observed concordance to the reference genotypes. The 20 selected MDA products were subjected to amplification-free sequencing library preparation and sequencing on eight lanes of Illumina HiSeq X within the Sanger Pipelines.

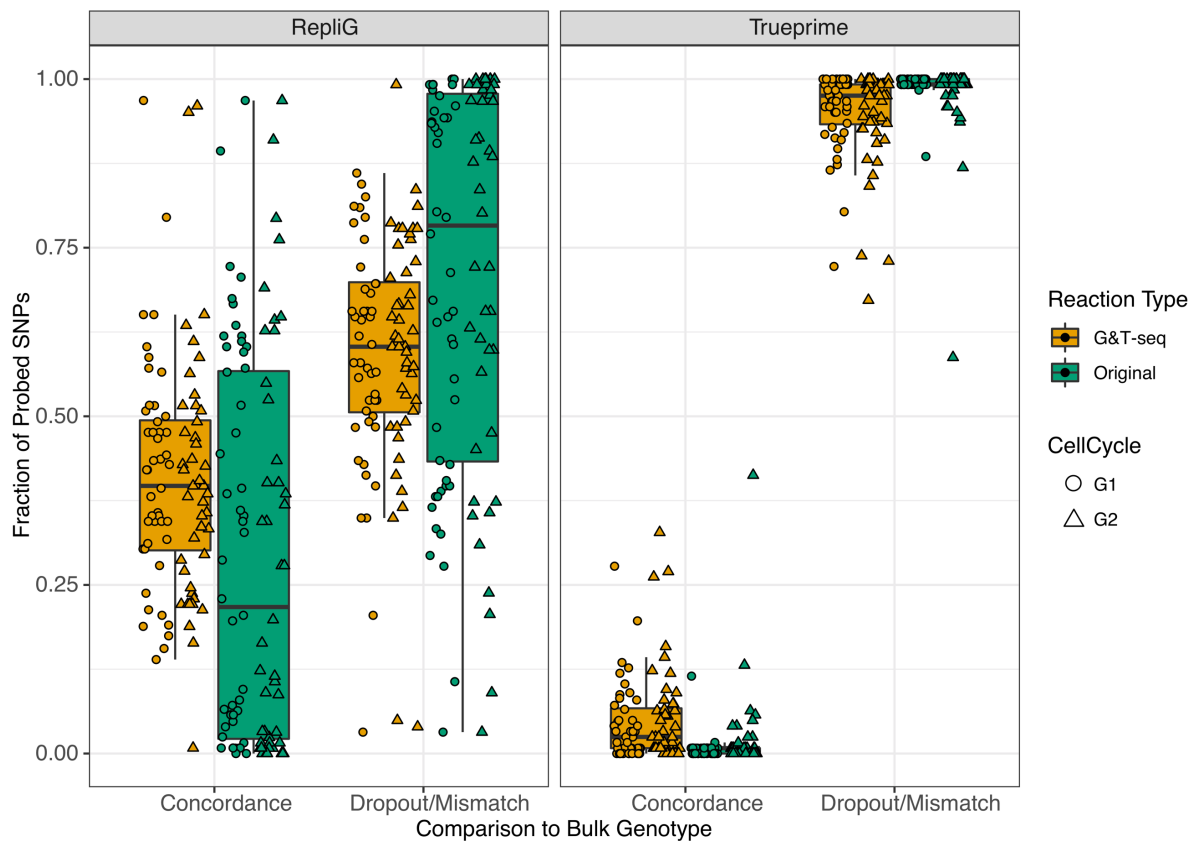


Figure 12: GbS Results for RepliG and Trueprime MDA Products. Genotypes called in the MDA products were compared to reference genotypes called in bulk sequencing data of the parental cell line. MDA genotypes were classified as reference-concordant or as displaying a mismatched or insufficiently covered locus for the 126 polymorphic sites probed by GbS. An average of 41% of the probed loci were called as reference concordant for the G&T-seq RepliG version, while the average for the corresponding original version was 30%, mainly due to substantially more cells with nearly no concordant calls. For Trueprime, most MDA products displayed less than 10% reference-concordance and even the most successful amplifications only 33% and 41% for the G&T-seq and the original version, respectively.

Coverage Statistics are Comparable for G&T-seq and Original RepliG Products

WGS data from the 20 selected MDA products was aligned to the human genome build GRCh37 using bwa mem as described in the methods section. When comparing mapping statistics, no difference between the G&T-seq and the original RepliG version or between the cell cycle states could be observed (Figure 13). For the G&T-seq version, an average of 92.4% (SD = 0.6%) of reads were successfully mapped, 7.3% (SD = 0.6%) were marked as duplicates and 0.3% (SD = 0.1%) were not aligned. Similar values were obtained for the original version with 92.3% (SD = 0.6%) mapped, 7.3% (SD = 0.6%) duplicated and 0.4% (SD = 0.1%) unmapped reads. Consequently, the obtained depth was also comparable with an average coverage of 13x (SD = 0.8) and 12.8x (SD = 0.6) for the G&T-seq and the original version, respectively. The even coverage obtained from the pooled sequencing indicated a comparable library complexity across all selected conditions.

One of the main advantages of MDA is the greater breadth of genome coverage compared to other WGA approaches^{105,176}. The cumulative fraction of the genome covered at depths between 0 – 50x coverage is displayed in Figure 14. All four 50-cell controls achieved a total genome coverage of 88% with 32% and 33% having a coverage of at least 15x for the G&T-seq version and 31% and 37% for the original version for the G1 and G2 cell cycle state, respectively. For single cells, no significant difference in the average breadth of coverage could be detected between the cell cycle states for single cells at any of the coverage thresholds considered. When comparing the MDA products for single cells from the original RepliG version, it was found that they achieved a slightly higher breadth of coverage of 69% (SD = 13%) compared to the G&T-seq RepliG products with 59% (SD = 14%). Notably, this difference is only maintained at low depth and for a coverage of at least 15x, both the original and the G&T-seq version cover 23% (SD = 4%) of the genome. As SNV calling is only reliable at greater sequencing depths, no meaningful difference could be detected between the original and the G&T-seq version.

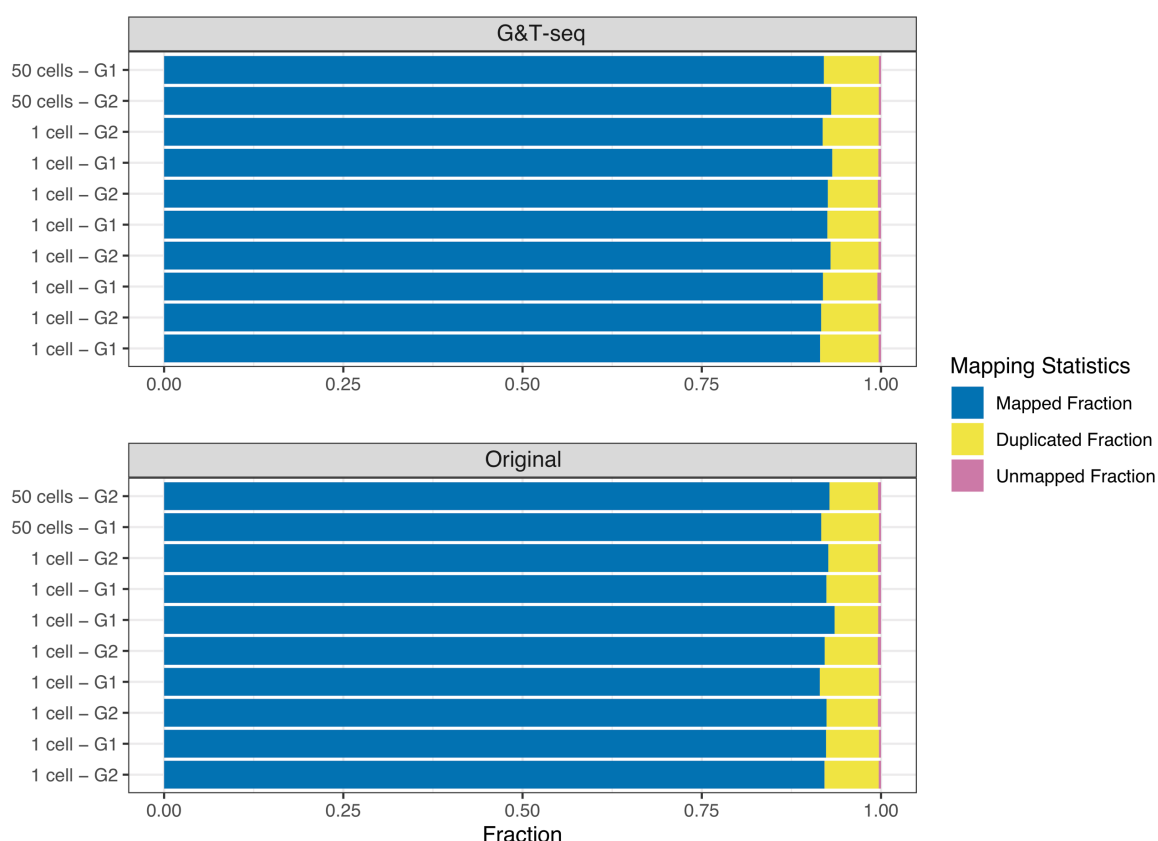


Figure 13: Mapping Statistics for WGS Data from RepliG MDA products. WGS data from twenty selected RepliG MDA products was aligned to GRCh37. The products were evenly selected for the G&T-seq and the original RepliG version as well as for the G1 and G2 cell cycle stage and comprise all four 50-cell controls. A similar fraction of 92% of the reads were successfully mapped with a similar percentage of duplicated and unmapped reads, when the original and G&T-seq version were compared. Furthermore, no difference could be observed between 50-cell controls and single cells or between the cell cycle states.

Besides breadth of coverage, the uniformity of coverage is an essential quality control parameter for single cell WGS data to control for the evenness of amplification. The inequality of the coverage distribution was computed and corresponding Lorenz Curves are shown in Figure 15. In addition to the greater breadth of coverage, the 50-cell controls also display a more uniform coverage distribution compared to single cells but are similar between the original and the G&T-seq version as well as between cell cycle states. For single cells, the Gini coefficients indicate slightly greater inequality for the G&T-seq version with an average Gini coefficient of 0.78 (SD = 0.08) compared to 0.74 (SD = 0.08) for the original version. However, this difference is not found to be statistically significant (Wilcoxon rank sum test, $p = 0.32$). Moreover, no difference between the G1 and G2 cell cycle states could be observed.

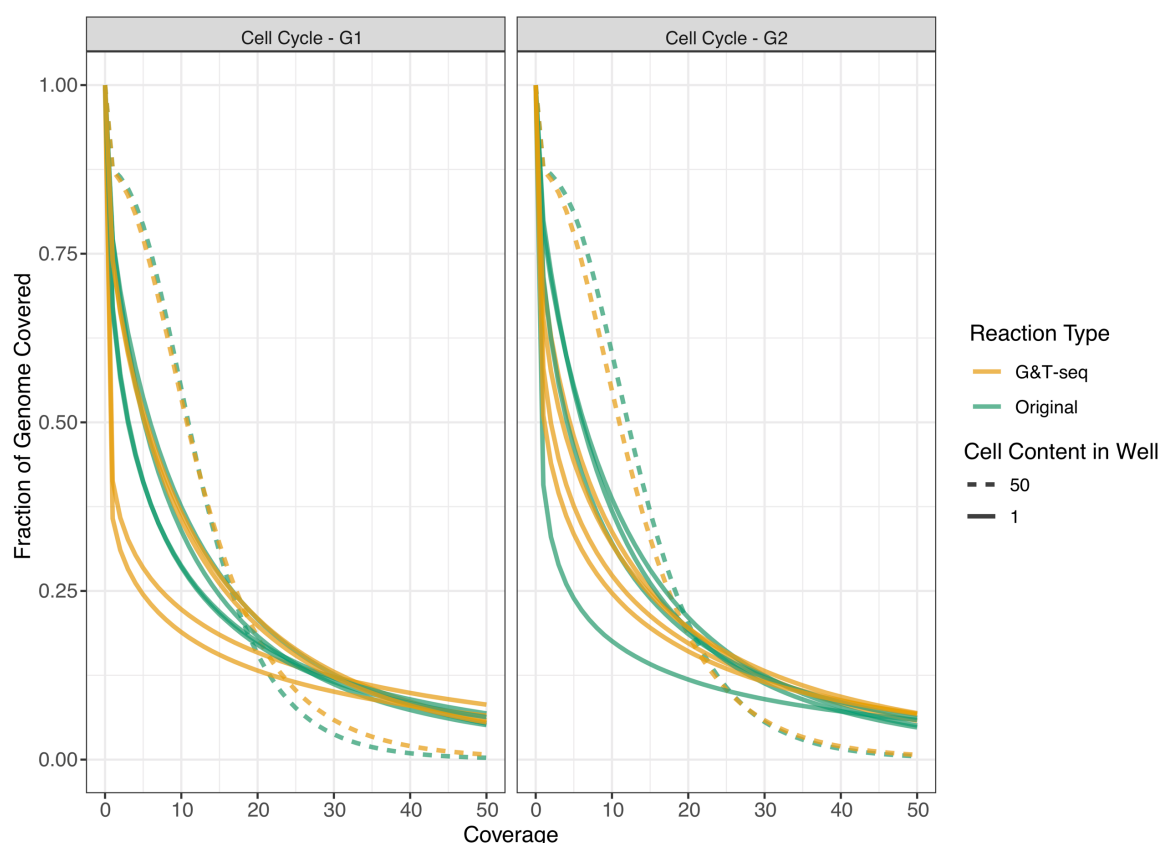


Figure 14: Cumulative Fraction of Genome Covered for WGS Data from RepliG MDA Products. The cumulative fraction of the genome covered at a minimum of depths between 0 – 50x was computed and illustrated above. The 50-cell controls show greater breadth of coverage than the corresponding single cells. While the average for the total breadth of coverage is slightly higher for the original RepliG version with 69% compared to 59% for the G&T-seq version, this difference was not maintained at coverages greater than 15x. No difference in the breadth of coverage could be detected between cell cycle states.

In summary, the WGS from G&T-seq and original RepliG amplification displayed no significant differences regarding the average depth, breadth of coverage at depths relevant for SNV analyses or uniformity of coverage. Collectively, this indicates sequencing libraries of similar quality can be derived within G&T-seq compared to the original RepliG version and the precipitation, transfer and MDA in presence of partially bead-bound DNA does not lead to a systematic loss of DNA molecules.

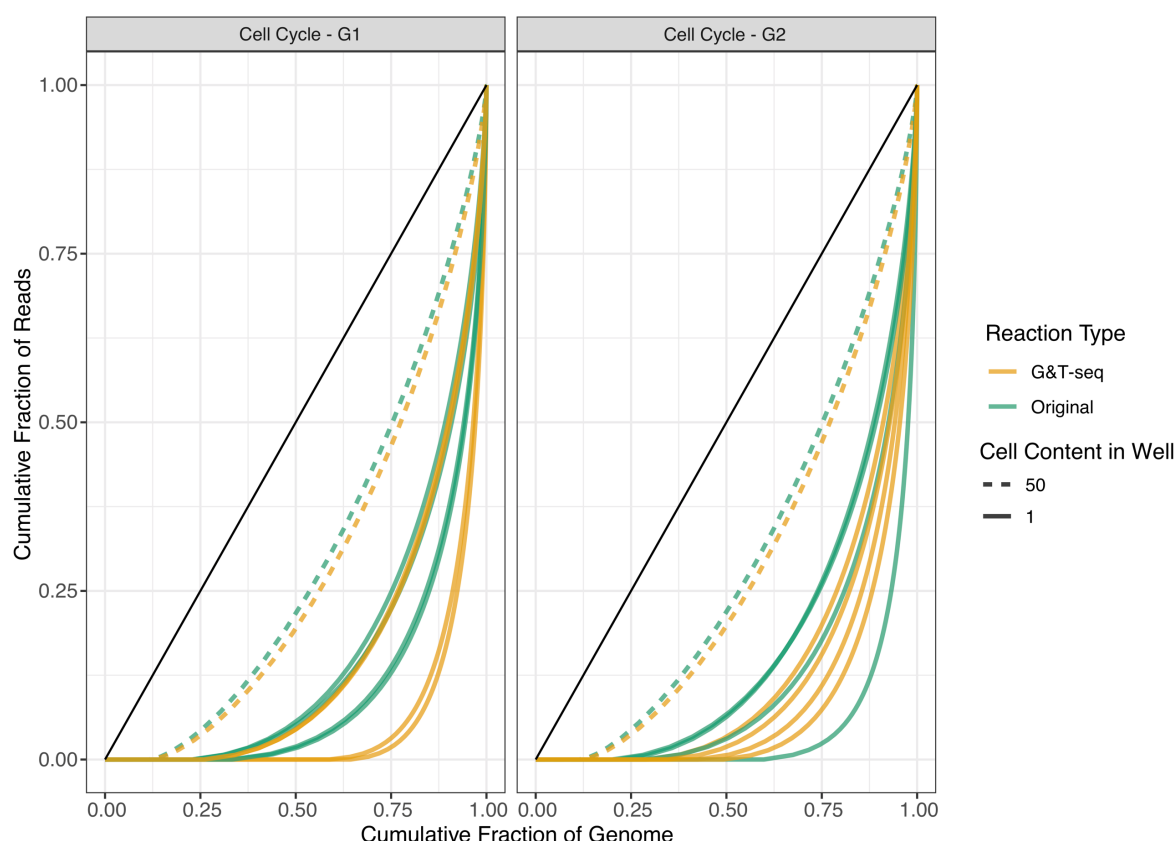


Figure 15: Uniformity of Genome Coverage for WGS Data from RepliG MDA Products. Inequality of coverage distribution was computed and illustrated as Lorenz Curves. The straight black line indicates perfectly even coverage. Single cells displayed greater inequality of coverage than the 50-cell controls. The Gini coefficients from the Lorenz curves of single cells did not differ significantly between the original and the G&T-seq version, suggesting similar uniformity of coverage. Furthermore, no difference between the G1 and G2 cell cycle state could be observed.

Mutational Burden and Spectra are not Altered in G&T-seq RepliG

After establishing that sequencing libraries of similar quality could be derived for G&T-seq compared to the original RepliG version, the impact on the false positive rate, mutational burden and spectrum was considered. While it is known that MDA introduces a substantial level of technical artefacts, it was important to verify that no additional level of noise is introduced within the G&T-seq workflow. Given the limited sequencing depth of the WGS from MDA products of about 13x on average, a genotyping approach on loci with known reference genotypes was performed as described in the methods section to estimate the false positive and ADO rate. A total of 1,363,033 homozygous and 515,780 heterozygous reference loci was considered and genotyped in the MDA-derived WGS.

From the total of 1,878,813 known reference sites, an average of 911,652 (SD = 325,034) and 835,111 (SD = 330,671) sites could be called in the MDA-derived WGS for the original and the G&T-seq version, respectively. To avoid any differences in the rates of false positives that are based on differences in the ability to call genotypes, any further calculations were based on the sum of successfully genotyped loci per cell. Only considering successfully genotyped sites, the fraction of false positive calls was 2.6% (SD = 0.9%) and 3% (SD = 1.2%) for the original and the G&T-seq version, respectively (Figure 16). The average ADO rates at reference heterozygous sites were 34% (SD = 20.2%) for the original RepliG version and 44% (SD = 26.8%) for the G&T-seq version. Neither the difference in FP nor in ADO rates was statistically significant (Wilcoxon rank sum test, $p = 0.25$ for FP and $p = 0.13$ for ADO). Notably, the MDA products were picked across a quality spectrum as assessed by the amplicon-based GbS assay. The 50-cell controls displayed nearly identical FP and ADO rates between the G&T-seq and the original RepliG version. The most favourable single cells displayed ADO rates of 33% and 35% for the original and the G&T-seq version, respectively. This indicated that a thorough QC before WGS is necessary and potentially able to mitigate any minor differences that could be observed between the average FP and ADO rates from the original and the G&T-seq version.

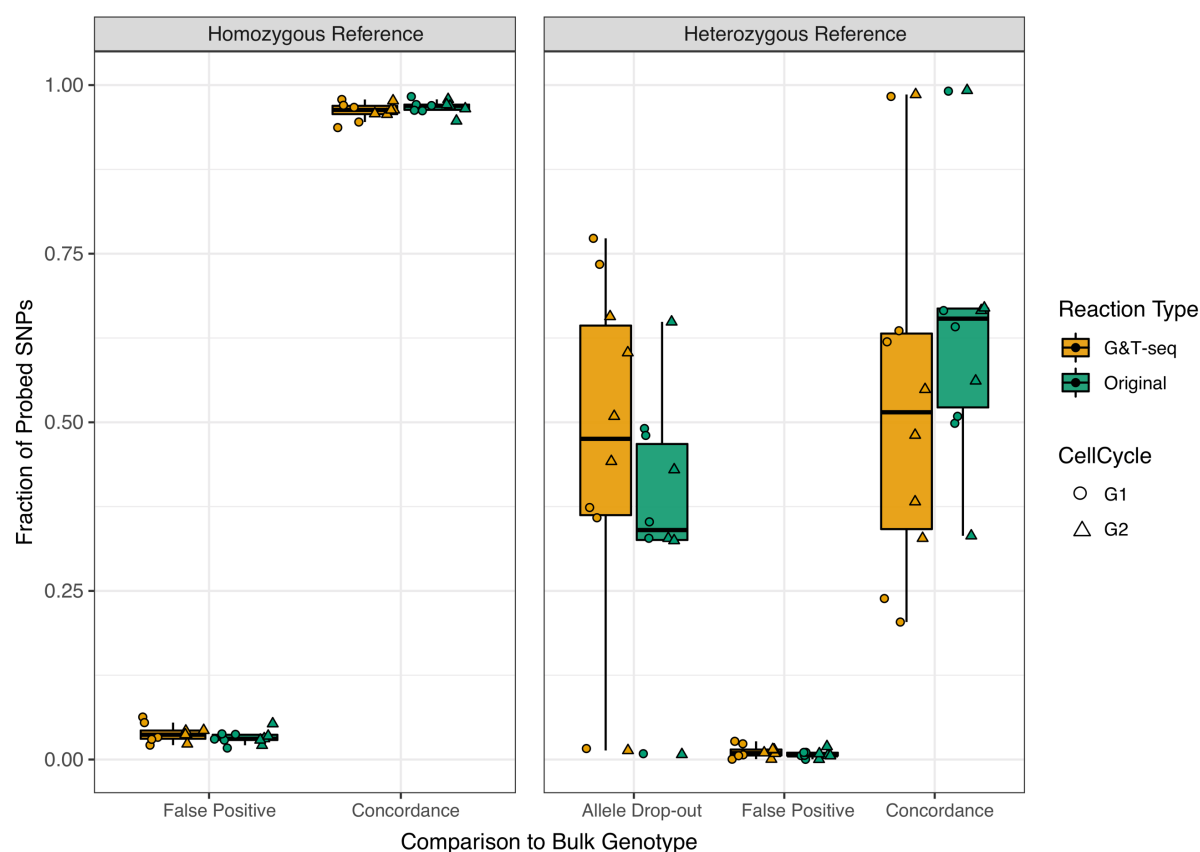


Figure 16: WGS-based Genotyping Results for RepliG MDA Products. A total of 1,878,813 loci with known reference genotypes derived from bulk sequencing of the parental cell line were genotyped within the WGS data of RepliG MDA products. The fraction presented is based on the total number of successful genotyping calls per MDA product to avoid any impact of differential coverage between MDA products. In total, similar average fractions of 2.6% and 3% were classified as false positive genotype calls for MDA products from the original and the G&T-seq version, respectively. For heterozygous sites, an average ADO rate of 34% and 44% was observed for MDA products from the original and the G&T-seq version, respectively. Notably, the most favourable single cells achieved a more comparable ADO rate between the two versions with 33% for the original compared to 35% for the G&T-seq version.

To derive mutational spectra, SNVs were called on a whole-genome level using the single cell-specific algorithm SCCaller, that is trying to avoid MDA artefacts through estimation of local amplification bias and adaptive filtering based on this bias. A more detailed explanation is given in the methods section. After artefact filtering and removal of common database SNPs, an average of 47,396 (SD = 15,418) mutations per MDA product were called for the original version and 47,592 (SD = 12,107) mutations per MDA product for the G&T-seq version. The mutational spectrum was derived as described in the methods section and is shown in Figure 17. The mutational spectra

were highly correlated between the G&T-seq and the original version (Pearson's $\rho = 0.99$, $p < 2.2e-16$), indicating that no additional artefacts are introduced through any of the additional G&T-seq processing steps.

In conclusion, the observed FP and ADO rates were comparable between G&T-seq and the original RepliG version and nearly identical for the most favourable single cell and 50-cell controls. Furthermore, the genome-wide burden as well as the mutational spectra were practically indistinguishable, strongly suggesting that genomic data derived from G&T-seq is comparable to traditional single-layer methods.

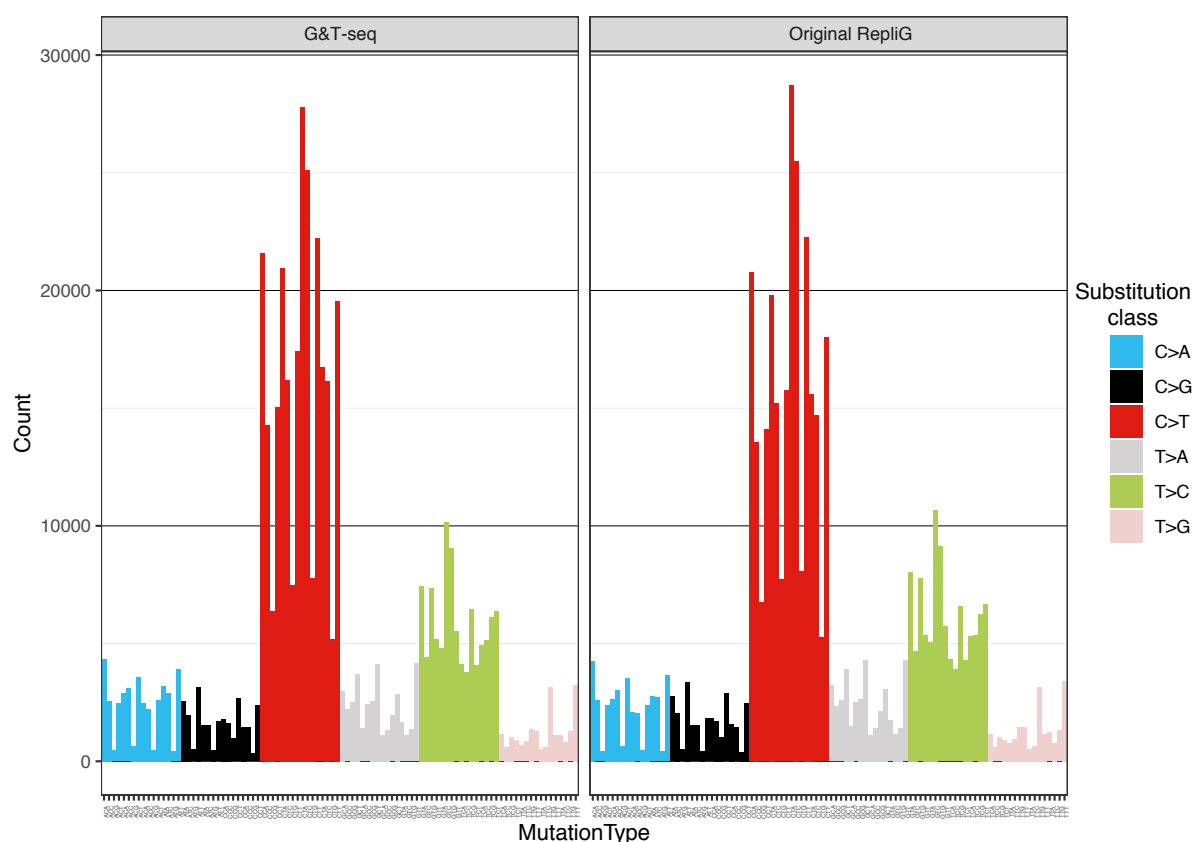


Figure 17: Base Substitution Mutational Spectra for Genome-wide SNV Calls. Genome-wide SNVs were called using the single cell-specific algorithm SCCaller. Base substitutions were classified according to the pyrimidine base change of the affected Watson-Crick base pair and further distinguished by their 5'- and 3'- flanking base context. The mutational spectra for the G&T-seq and the original RepliG version were nearly identical, indicating that the G&T-seq processing steps did not create additional artefacts.

GbS Represents an Effective Tool for MDA Quality Control Before WGS

The amplicon-based assay GbS was chosen as tool for MDA quality control before WGS as previous work demonstrated that one qPCR per chromosome could be used to distinguish between failed and successful MDA reactions ²²⁹. The GbS assay is performed using multi-well plate formats and liquid handling robots, which substantially reduces the manual labour per tested MDA product, while providing genotype-level information in 126 loci across all human chromosomes compared to only a binary classification of successful amplification for all autosomes in the previously published approach ²²⁹. As explained in the section above describing the GbS results, MDA products were chosen for WGS across a range of GbS-indicated quality for all considered conditions. While the original and the G&T-seq version of RepliG were shown to be comparable for all the considered metrics, there was always substantial deviation within every condition. To consider how well the WGS quality compared to the quality indicated by GbS, the percentage of reference-concordant genotypes per MDA product was chosen as numeric metric for GbS quality and compared to several WGS-derived quality criteria.

For all WGS-derived quality criteria considered, the GbS-indicated quality score displayed a strong correlation and the general pattern was independent of the cell cycle state and application of the G&T-seq or original RepliG version. Consequently, ordinary least squares regression (OLS) was performed for the complete data set to evaluate the statistical significance of GbS and the respective WGS quality scores (Figure 18). As SNV calling is only reliable at moderate to high sequencing depth, the fraction of the genome covered at a depth of at least 15x was chosen to represent the meaningful breadth of coverage, which displayed a positive correlation with the fraction of reference-concordant genotypes per MDA product (adjusted $R^2 = 0.84$, $p = 1.03e-8$; Figure 18A). To assess the how well uniformity of coverage is predicted by GbS, the Gini coefficients from WGS of MDA products were used. High Gini coefficients indicate greater inequality and a significant negative correlation with the GbS-indicated quality could be observed (adjusted $R^2 = 0.9$, $p = 1.06e-10$; Figure 18B). ADO as assessed during the genotyping analysis of WGS from MDA products was compared to the fraction of reference-concordant genotype calls from the GbS results as another critical parameter of WGS quality. A greater reference-concordance during GbS strongly correlated with low ADO values in WGS data, suggesting a less biased

amplification in MDA products with higher GbS quality scores (adjusted $R^2 = 0.84$, $p = 9.69\text{e-}9$; Figure 18C). Finally, the genome-wide burden of mutations also displayed a significantly negative correlation with the GbS quality score (adjusted $R^2 = 0.35$, $p = 3.53\text{e-}3$; Figure 18D). As MDA is known to produce a high level of artefacts, a reduced overall burden of mutations across cells with similar genetic background, suggested a lower level of technical noise in cells with greater GbS quality scores.

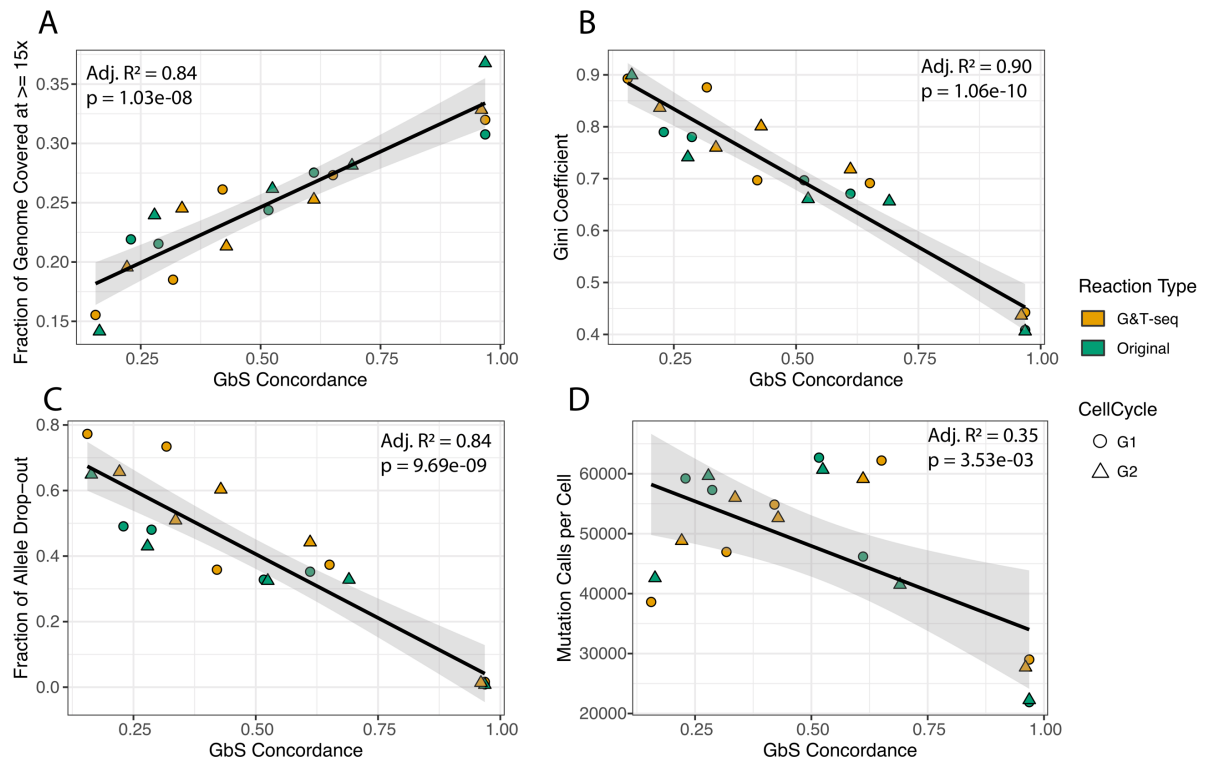


Figure 18: Correlation of GbS-indicated Quality with WGS-derived Quality Metrics. Linear regression was performed between the four displayed WGS-derived quality metrics and the fraction of reference-concordant genotype calls as assessed by the GbS assay before sequencing. Strong correlations were observed for all considered metrics, which suggested that GbS represents an effective tool to screen for high-quality MDA products. **(A)** Breadth of coverage displayed as fraction of the genome with a depth of at least 15x. **(B)** Uniformity of coverage as measured by the Gini coefficient. **(C)** ADO rates derived from genotyping in the WGS data. **(D)** The genome-wide burden of SNV calls.

Discussion

Whole-genome amplification is a crucial processing step for single cell sequencing and several comparisons have been published to identify the most suitable chemistries for their particular experimental setting and downstream analyses ^{105,106,194-200}. While a general consensus appears that hybrid methods are superior for CNV analyses and MDA-based approaches are more suitable for SNV analyses, the final results vary given the experimental settings, metrics used and chemistry kits considered. As G&T-seq includes several unique processing steps, a comparison of three MDA chemistry kits was performed to assess their suitability within this workflow and to ensure that the additional processing steps do not impair the WGS quality. Moreover, the semi-automated amplicon-based assay GbS was evaluated as cost-effective and scalable means to assess the MDA amplification quality before expensive WGS.

For each of the three selected chemistries - GenomiPhi, Trueprime and RepliG – a total of 188 MDA reactions was performed, evenly split between cells that had undergone the additional processing steps necessary for G&T-seq as well as cells that were directly subjected to the MDA reactions according to the respective manufacturer's instructions. The MDA reactions were assessed according to their yield (Figure 11) as well as their concordance with reference genotypes derived from bulk sequencing data from the same parental cell line using GbS (Figure 12).

GenomiPhi amplifications failed to reach the required minimum of 500 ng to perform amplification-free sequencing library generation within the Sanger Pipelines for cells that had undergone the additional G&T-seq processing steps. Notably, standard GenomiPhi amplifications performed directly on single cells exceeded this threshold four-fold on average. The additional processing steps within G&T-seq require the MDA reaction to be performed in the presence of potentially partially bead-bound DNA and trace residues of ethanol. Additionally, cells have to be lysed within a different lysis buffer for G&T-seq compared to the manufacturer's recommendations for GenomiPhi. These differences are potentially accountable for the lower efficiency of G&T-seq GenomiPhi reactions compared to the original version. Notably, the difference in yield between the G&T-seq and the original version was only observed for GenomiPhi but

neither for Trueprime nor for RepliG (Figure 11). Thus, the specific and non-disclosed reaction conditions of the GenomiPhi kit were decided not to be suitable for G&T-seq. While Trueprime reactions yielded 1.4 µg on average and thus, were deemed to be in a suitable range for amplification-free sequencing library preparations, nearly all RepliG reactions generated at least 10 µg up to over 30 µg of DNA. According to the manufacturer's instructions, the MDA for Trueprime should be performed for only 3 h in contrast to the 8 h recommended for RepliG. As the product yield scales with the duration of amplification and time-limited MDA has been demonstrated to generate less technical artefacts, the long amplification time recommended for RepliG could potentially be reduced for future experiments to restrict technical noise while still creating enough product for amplification-free sequencing library preparation ^{205,229}.

The MDA reactions from Trueprime and RepliG were subjected to GbS to assess their genotype concordance with bulk sequencing data from the same parental cell line. As a previous publication demonstrated that qPCR-based methods could be used to verify the successful amplification per chromosome, I wanted to extend on this idea ²²⁹. Indeed, great variability within MDA product quality was indicated for RepliG reactions with an average genotype concordance of 33% for the original and 41% for the G&T-seq version while Trueprime reactions only showed minimal concordance with the bulk sequencing data (Figure 12). Motivated by the GbS results, twenty RepliG products were selected for WGS across a range of GbS quality scores, evenly split between the original and G&T-seq RepliG version and the G1 and G2 cell cycle state of the originally sorted cells. As GbS was shown to be an effective predictor of WGS quality later on (Figure 18), it can be assumed that WGS derived from Trueprime MDA products would have been of inferior quality compared to the RepliG products. Another MDA comparison that included Trueprime and RepliG evaluated their WGA quality based on amplicon-based sequencing of over 3400 loci. Similar to the results presented in this thesis, most Trueprime products were considered failed reactions due to their poor amplicon recovery rate, while RepliG was considered to be the most suitable for SNV analyses of the seven WGA chemistries considered ¹⁹⁷. Another MDA comparison considering Trueprime also concluded that Trueprime products are not suitable for CNV analyses due to high locus dropout rates and substantially biased amplification ¹⁹⁹. The main difference between Trueprime and RepliG is the usage of a DNA primase within Trueprime compared to random oligonucleotide priming within

RepliG. While the DNA primase was postulated to result in more evenly amplified genomes compared to the random priming in traditional MDA approaches, only the original publication on Trueprime amplification supports this claim so far ²²⁸.

The RepliG products selected for WGS were assessed for general genome sequencing quality criteria including their read mapping statistics, breadth and uniformity of coverage (Figure 13-15). These criteria were compared between 50-cell controls and single cells, the G1 and G2 cell cycle states and most importantly, between the G&T-seq and the original RepliG version. In general, the 50-cell controls displayed a greater breadth and a more even genome coverage compared to single cells as expected. However, no substantial or significant differences could be found between the G&T-seq and the original RepliG version for any of these metrics (Figure 13-15). Moreover, the results were also comparable between cell cycle states, not supporting a claim from previously published work that MDA amplifications are more successful when starting from material of cells in G2 phase ²²⁹.

Furthermore, quality criteria more specific to SNV analyses were assessed. The ADO and false positive rates were assessed through comparison to reference genotypes from bulk sequencing data from the same parental cell line (Figure 16). While the false positive rates were very similar between the G&T-seq and the original version with 2.6% and 3%, respectively, the average ADO rates were about 10% higher for products generated from G&T-seq samples compared to the original version. However, the most favourable single cell products achieved similar ADO values of 33% and 35% for the original and the G&T-seq version, respectively. As the MDA products were selected across a range of GbS quality values and these were demonstrated to be able to predict ADO rates of WGS data (Figure 18C), this indicated that the difference in ADO rates is not necessarily systematic and can be mitigated by careful quality control before WGS. Moreover, the genome-wide burden of base substitutions was assessed using SNV calls from a dedicated single cell algorithm and the resulting mutational spectra were compared between the original and the G&T-seq version. Despite the additional processing steps necessary for G&T-seq, MDA products had comparable burdens of 47,396 mutations for the original and 47,592 mutations for the G&T-seq version. As their corresponding mutational spectra were also nearly identical (Figure 17), this indicates that the processing steps necessary for

G&T-seq do not create substantial additional technical noise and the obtained genome sequencing data is of similar quality compared to conventional approaches. Notably, the cell cycle state did not have any impact on the SNV analyses-specific quality criteria. This is an important finding in the context of post-mitotic tissues that have no or only very few cells in the G2 cell cycle state as these can be analysed using a G&T-seq RepliG approach without any considerations beyond the general challenges of calling SNVs in MDA-amplified sequencing data ^{51,104,176,205,229}.

In summary, G&T-seq was successfully shown not to interfere with RepliG MDA reactions and the corresponding genomics data is suitable for SNV analyses independent of the cell cycle state of the tissue of interest. The insufficient yield of GenomiPhi reactions within G&T-seq and the general lack of high-quality Trueprime products, highlight the need for thorough WGA consideration within new experimental settings. Moreover, GbS could be identified as powerful predictor of WGS quality, enabling scalable and cost-efficient screening of MDA products before WGS. Collectively, these findings support the extension of G&T-seq to SNV analyses and identified effective methods for the necessary WGA and corresponding quality control.

A Multiomics Approach to Define Mutational Spectra in Single Neurons

Introduction

This chapter describes the application of G&T-seq to cortical neurons to identify somatic base substitutions. Previous studies published just prior to and during the time of this analysis have used MDA amplified data to explore the mutational burden and signatures in human neurons. However, they did not have access to transcriptomic data of the same single cell ^{102,103}. While the more recent study acknowledged a high burden of MDA artefacts in the data from the first publication and addressed this with computational filtering methods, an integrated DNA-and-RNA approach enables the confirmation of somatic variants in a molecular layer that is independent from MDA amplification and the corresponding artefacts.

Human cortical development as well as the relevance and burden of somatic mutations in normal brain and neurodegeneration are briefly summarised as background for this chapter.

Human Cortical Development and Neurogenesis

Human brain development starts with the formation of the neural tube around three weeks after conception. Roughly one week later, the three vesicles called prosencephalon, mesencephalon and rhombencephalon have budded of the neural tube, which form the forebrain, the midbrain and the hindbrain, respectively. During the course of another week, the prosencephalon divides into the diencephalon, which amongst other structures will develop into the thalamus and hypothalamus, and the telencephalon, which becomes the cerebral cortex (reviewed in ²³⁰). The cortex accounts for over 80% of the human brain mass and is involved in several higher functions including processing of language and most sensory stimuli as well as the regulation of voluntary movement ²³¹.

Following the initial compartmentalisation, cortical development proceeds with a phase of rapid neuronal proliferation and migration, followed and partially overlapped

by series of apoptosis, the generation of synapses and axon myelination ²³⁰. Generation and migration of neurons is different for the two main neuronal subtypes of the cortex. The excitatory, glutamatergic projection neurons are generated in the dorsal telencephalon and migrate mostly radially over short distances, while the inhibitory, GABAergic interneurons primarily derive from the ventral telencephalon and have to migrate tangentially over longer distances (Figure 19A) ²³².

At the onset of neurogenesis in the dorsal telencephalon, the founder population of neuroepithelial cells (NECs) turn into radial glial cells that are associated with the apical membrane of the ventricular zone (VZ) and span the entire thickness of the developing cortex. Radial glial cells are stem cells that give rise to excitatory neurons and eventually differentiate into astrocytes. Besides symmetric division for self-renewal, radial glial cells can asymmetrically divide to directly give rise to neurons or to intermediate progenitor cells (IPCs). These IPCs migrate into the subventricular zone (SVZ), which is directly above the VZ, where they significantly contribute to neurogenesis in humans. The excitatory neurons generated in these regions attach to the radial glial cells and migrate alongside their processes to populate the eventually six-layered cortex. The six layers of the fully developed neocortex differ with respect to the relative fraction and subtype of excitatory neurons and are involved in different neurocircuits (reviewed in ²³²⁻²³⁴). Cortical GABAergic interneurons can be found in all of the six layers as well with a similarly variable fraction of inhibitory subtypes in different layers. While there is some evidence for GABAergic neurogenesis directly within the developing human neocortex, the vast majority of interneurons derives from the ganglionic eminence, which is a transient embryonic structure of the ventral telencephalon. Therefore, new GABAergic neurons first have to migrate tangentially into the neocortex before they are integrated into the cortical layers after their attachment to radial glial cells (Figure 19B). Although the tangential migration across a large distance compared to the short radial migration, lacks the physical guidance of radial glial or similar cells, it is highly specific as clonally related interneurons were demonstrated to cluster spatially in the fully developed neocortex (reviewed in ²³²⁻²³⁴).

Notably, the fully developed human cerebral cortex comprises around 16 billion neurons with the majority being generated over a short time span of four to six weeks of embryonic development ^{230,231,235}. While adult human stem cells can give rise to

neurons in the olfactory bulb and the hippocampus, cerebral neurons are post-mitotic cells that persist for a whole human lifespan after embryonic proliferation eventually subsides around 26 weeks after conception ^{235,236}.

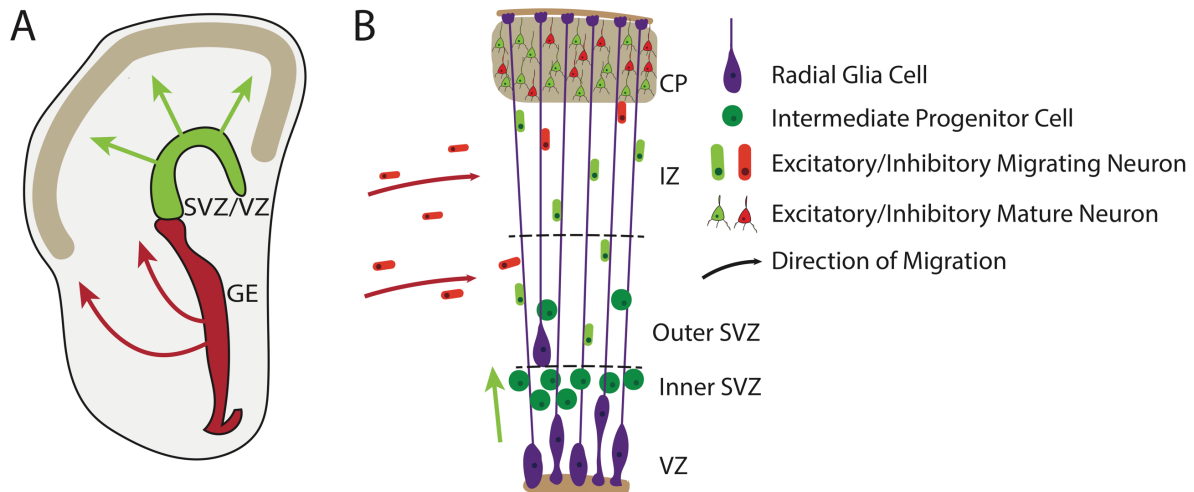


Figure 19: Neurogenesis and Neuronal Migration. SVZ: Subventricular Zone. VZ: Ventricular Zone. GE: Ganglionic Eminence. IZ: Intermediate Zone. CP: Cortical Plate. **(A)** Cortical excitatory and inhibitory neurons originate from morphologically distinct regions of the developing brain. Excitatory neurons are generated in the ventricular and subventricular zone while the majority of inhibitory neurons derives from the ganglionic eminence. **(B)** Excitatory and inhibitory neurons are the main neuronal cell types in the human cortex. Excitatory neurons migrate primarily radially along radial glial cells. Inhibitory neurons first migrate tangentially without physical guidance before attaching to radial glial cells. Both types of neurons are interconnected in neural circuits throughout all cortical layers.

Somatic Mutation in Normal Brain and in Neurodegeneration

The concept of somatic mosaicism is discussed in the general introduction to this thesis ^{45,46}. Similar to other tissues, nearly all cells in the brain will carry clonal and private mutations that were caused by various endogenous and exogenous factors including replication, transcription or mutagen exposure, resulting in a high degree of intra-tissue heterogeneity ^{13-15,237}. In recent years, especially single cell sequencing methods provided the opportunity for detailed exploration of the somatic mosaicism in the human brain. Substantial differences have been observed between the fetal and the adult brain with much higher rates of mutation during earlier stages of development ^{87,95-97,102,238}. Furthermore, an age-related effect on SNVs and CNVs was demonstrated even within post-mitotic neurons of the adult brain ^{102,239}.

In more detail, whole-chromosome aneuploidy rates between 30-40% have been indicated by single cell sequencing and fluorescence *in-situ* approaches in the fetal brain, while a prevalence of only 2% was found in the adult brain ^{95-97,238}. L1 retrotransposition is another process that is highly active during embryonic development of the brain ^{240,241}. Several studies have analysed the corresponding prevalence in the adult brain with variable estimates between less than one or more than ten retrotransposition events per neuron on average ^{98,101,242}. Notably, a re-analysis of the study claiming ubiquitous L1 retrotransposition in the human hippocampus, challenged the original author's high estimate of more than ten events per neuron and aligned their data with the consensus of less than one retrotransposition per neuron. The re-analysis identified several sources of artefacts and highlights the importance of careful examination of single cell sequencing-derived data ¹⁰⁰. In contrast to low levels of aneuploidy and retrotransposition, a substantial fraction of 5% to 30% of all neurons in neurotypical individuals carried at least one mega-base scaled CNV with an average of 63 genes being affected per alteration ^{94,96,97,239,243}. Interestingly, the relative percentage of neocortical neurons with CNVs declines with age. As neocortical neurons enter a post-mitotic state upon differentiation during embryonic development, the decline in CNV prevalence suggests an increased vulnerability to age-associated atrophy for neurons with particular mutations ²³⁹.

The most abundant type of somatic mutations in the human brain are SNVs ^{87,102,103}. Several hundred base substitutions could be already detected in single cell-derived colonies from neural progenitor cells that were sampled from post-mortem human foetuses 15 to 21 weeks after conception. Remarkably, SNVs accumulated at variable rates during different developmental phases with around one mutation per cell division for early cleavage stages and over five manifested base substitutions per day during the highly proliferative stage of neurogenesis. The observed mutation rates would result in around 1,000 somatic SNVs per neuron at the time of birth ⁸⁷. Two additional studies have investigated somatic SNVs in individual human neurons using single cell genome sequencing of MDA products ^{102,103}. The first study found an average of around 1,500 SNVs per neuron across three different individuals with relative enrichment in actively transcribed regions in neurons. The second study from the same authors acknowledged the potential existence of MDA artefacts in their first

publication and implemented a computational approach to distinguish true variants from technical artefacts using phasing of candidate sites with germline SNPs ^{102,206}. Fifteen neurotypical individuals spanning an age range of over 80 years were investigated with less than 1,000 SNVs for young individuals and between 3,000 – 6,000 base substitutions for donors at around 80 years of age. Hence, an age-dependent pattern of SNV accumulation could be detected with an average of 23 and 40 mutations per genome and year in post-mitotic neurons from the prefrontal cortex and the dentate gyrus, respectively ¹⁰².

The relevance of somatic mutations in the context of neurodegenerative diseases has initially been proposed based on the comparison of familial and sporadic cases of these diseases and simulations estimating the degree and impact of somatic mosaicism ^{244,245}. As mutations known from familial cases of diseases such as Alzheimer's and Parkinson's, could not be detected in the blood-derived DNA of sporadic cases, the mutations were suggested to be restricted to the central nervous system. Indeed, several somatic mutations including SNVs and CNVs in disease-relevant genes were demonstrated in human brain samples from affected donors by now (reviewed in ²⁴⁶). Additionally, post-mortem brain samples of patients with neurodegenerative disorders display an increased burden of somatic SNVs and distinct enrichment of L1 retrotransposition in intronic regions compared to non-pathological cases ^{102,247}.

Collectively, these findings demonstrate the substantial degree of somatic mosaicism in brains of neurotypical individuals and highlight the impact of particular somatic mutations in the context of neurodegeneration. Single cell sequencing is ideally suited to explore the mutational landscape in neurons due to their lack of proliferative capacity or a distinct clonal architecture, which would be needed for organoid or targeted microdissection approaches discussed in the general introduction of this thesis. Although previous studies have published important findings about the SNV burden and corresponding mutational signatures in post-mitotic neurons, they solely rely on MDA amplified data, which is known to comprise high levels of technical noise ¹⁷⁶. The substantial variability and partial disagreement in the estimated burden of CNVs and retrotransposition derived from single cell sequencing data, highlights the importance of complementary approaches in this still emerging field to accurately

determine mutational processes within single cells. Multiomics data from methods such as G&T-seq provides the unique opportunity to analyse information from different molecular layers from the same single cells. This does not only allow for a more detailed characterization based on analyses on the individual layers but also for an integrated approach. Since genomic information from G&T-seq was shown to be comparable to conventional MDA approaches and the RNA layer is independent of this WGA method, cortical neurons from five neurotypical individuals were subjected to G&T-seq and the corresponding first-time multiomics SNV calling approach is presented in this chapter.

Methods

Brain Sample Collection

Frozen post-mortem human brain tissues were obtained by the Voet research group at KU Leuven with appropriate ethical approval. Tissues were obtained in collaboration with Professor Bart Dermaut from the University of Ghent and the University of Lille and Professor Vincent Deramecourt from the Lille NeuroBank. The brain donors were not diagnosed with any neurodegenerative diseases and the brain tissues were assessed as histologically normal. Data from five donors was generated in this chapter and are referred to according to their naming convention at KU Leuven as Brain19, Brain20, Brain21, Brain22 and Brain23. The data set covered an age range of around three decades, ranging between 40 and just over 70 years of age.

Isolation of Single Neuronal Nuclei

Isolation of neuronal nuclei was performed by Sarah Geurs at the KU Leuven. Briefly, frozen tissue blocks of around 0.25 cm³ from the prefrontal cortex were manually homogenised using a pre-cooled cell lysis buffer and a tissue grinder. The homogenate was layered on top of a sucrose density gradient and nuclei were separated from cellular debris via ultracentrifugation at 4 °C. The nuclear pellet was resuspended and stained using DAPI and the neuron-specific antibody NeuN ²⁴⁸. Individual neuronal nuclei were isolated via FACS. Doublet discrimination was performed using a Forward-Scatter height versus Forward-Scatter area approach and neuronal nuclei were discriminated by positive NeuN staining. Neuronal nuclei were sorted into 96-well plates containing 2.5 µL Buffer RLT Plus from Qiagen supplemented with 1:16M dilution of External RNA Controls Consortium (ERCC) spike-in controls ²⁴⁹. A full 96-well plate was sorted for Brain23 and half a plate each for the remaining four donors. Therefore, Brain19 and Brain20 as well as Brain21 and Brain22 were processed in common batches. The 96-well plates were briefly spun down and stored at -80 °C before further processing.

G&T-seq for Neuronal Nuclei.

All processing steps for G&T-seq were done by Sarah Geurs at the KU Leuven. Separation of mRNA and gDNA as well as mRNA processing were performed according to the original G&T-seq publications with MDA amplifications of the gDNA

layer as described in a previous chapter ^{177,178}. Briefly, polyA-tailed RNA was separated from gDNA using magnetic beads with oligo-dT primers. The gDNA was transferred into a new 96-well plate, while the mRNA was converted to cDNA using the Smart-seq2 method to derive full-length transcripts ^{250,251}. The gDNA was amplified using the RepliG protocol with restricted MDA duration of 3 h. The obtained cDNA and amplified gDNA was sent to the Wellcome Sanger Institute for GbS, library preparation and sequencing.

DNA Extraction for Bulk Sequencing Controls

In addition to isolation of individual neuronal nuclei from the frontal cortex for G&T-seq, DNA extractions from frozen brain tissue samples of the occipital lobe or the cerebellum of the same brain donors were performed. The obtained DNA was subjected to bulk genome sequencing and served as matched control during the analysis of the corresponding G&T-seq data. DNA extractions were performed by Sarah Geurs at the KU Leuven and sent to the Wellcome Sanger Institute for library preparation and sequencing.

Genotyping-by-Sequencing for MDA Product Quality Control

The lab work required for GbS was performed by Scott Goodwin and Naomi Park within the Sanger Pipelines. The amplicon-based assay as well as the subsequent genotype calling and comparison to the bulk reference genotypes were performed as described in the previous MDA implementation chapter.

Sequencing Library Preparation, Genome and Transcriptome Sequencing

Two different sequencing libraries were generated for the cDNA. The first round of cDNA libraries was created to obtain an overview of all cells that had been subjected to G&T-seq to determine the available neuronal subtypes in the data set. These libraries were generated in the Sanger Pipelines using the Nextera XT technology for RNA-seq from Illumina ²⁵². The pools of libraries per 96-well plate generated by the Nextera XT workflow were multiplexed in equimolar ratios and submitted for 125 bp paired-end sequencing on two lanes of the Illumina HiSeq 2000 using the v4 chemistry within the Sanger Pipelines.

After the selection of 50 successful MDA amplification products based on the GbS results from appropriate neuronal subtypes as indicated by the initial round of RNA sequencing, the corresponding original cDNA of these selected cells was submitted to the Sanger Pipelines for conventional Illumina sequencing library preparation and sequencing as an equimolar pool on two lanes of the Illumina HiSeq 2000 with v4 chemistry to yield 125 bp paired-end reads.

The 50 selected MDA products were submitted to the Sanger Pipelines for amplification-free sequencing library preparation using standard Illumina protocols ²¹⁹. The libraries for six to eight samples were multiplexed in equimolar ratios and sequenced on one lane of the Illumina HiSeq X platform per sample to generate 150 bp paired-end reads.

DNA extracted from corresponding bulk brain tissue was submitted to the Sanger Pipelines for conventional Illumina sequencing library preparation. These libraries were pooled in equimolar ratios and sequenced on one lane of the Illumina HiSeq X platform per sample to generate 150 bp paired-end reads.

Genome alignments for WGS and RNA-seq Data

RNA- and DNA-seq data was obtained as individual cram files per lane and per cell. The cram files were converted back into FAST format using biobambam ²²⁰. RNA- and DNA-seq data was concatenated separately per cell. The Nextera or conventional Illumina sequencing adapter, bases with phred-scaled quality scores less than 10 and unknown nucleotides were removed using TrimGalore and Cutadapt ²²¹. For RNA-seq data, trimmed reads with a length of less than 20 bp were discarded. RNA-seq data was aligned using STAR v2.5.2 with default parameters to a GRCh37 build supplemented with the ERCC sequences as additional chromosomes ²⁵³. Genome alignments of the WGS data to GRCh37 was performed with bwa mem and default parameters ²²². The genome alignments were sorted and indexed with samtools, STAR alignments only had to be indexed ²²³. Duplicates were removed for DNA-seq and flagged for RNA-seq data using Picard v1.6 ²²⁴.

Count Matrix Generation and Quality Control for RNA-seq Data

A count matrix from the STAR alignments was created with featureCounts against the Ensembl version 73 gene annotation ²⁵⁴. Only reads with a phred-scaled alignment score of greater than 30 were considered and fragment-based counts were summarised across Ensembl gene IDs. The count matrix was processed in R using the scater package ²⁵⁵. Features with less than five counts in less than two cells were removed to avoid spurious mapping to unexpressed genes. Mitochondrial genes as well as ERCCs were defined as control features and were used to filter for libraries that contained only limited biological information. Cells with less than 100,000 total counts or relative count fractions of at least 25% accounting for either mitochondrial genes or ERCCs were removed. Finally, a minimum of 1000 expressed endogenous features was required to consider an RNA-seq library for the subsequent cell type identification. The filtered count matrix was transformed into counts per million mapped reads (CPM) for size-normalisation and potential batch effects were considered using the Principal Component Analysis (PCA) from the scater package ²⁵⁵.

Computation of FPKM Values

FPKM values as measure for gene expression annotation were computed from STAR alignments against the Ensembl version 73 annotation using RSeQC ²⁵⁶.

Neuronal Subtype Identification

The filtered and size-normalised count matrix was loaded into SC3 for unsupervised consensus clustering to determine the cell types in the RNA-seq data set from NeuN-positive nuclei ²⁵⁷. SC3 performs parallel clustering using three different distance measures and two transformations and derives a consensus clustering from these individual clustering results. The clustering is k-means-based and a range of two to ten possible clusters was evaluated. The optimal number of clusters was evaluated by maximising the silhouette value of the consensus clustering. Unbiased marker gene identification was performed using SC3 ²⁵⁷. To determine marker genes, SC3 constructs a binary classifier for each gene based on the mean expression value for every cluster. Subsequently, SC3 computes the area under the receiver operating characteristic curve (AUROC) and a p-value for the robustness of the assignment using the Wilcoxon signed rank test. Here, marker genes were required to display an

AUROC > 0.8 and a Wilcoxon-derived p-value < 0.01. The obtained marker genes were compared to published marker genes from the literature to classify the neuronal subtypes that were present in the data set.

SNV Calling in RNA-seq Data

SNVs from RNA-seq data were called in the 50 neurons selected for the integrated analysis using GATK-v3.4 and the corresponding GATK Best Practices workflow for RNA-seq data²⁵⁸. Briefly, the novel splice junctions as discovered by STAR during the first alignment were added to the STAR index and the reads were re-aligned to the GRCh37 genome using the extended splice junction index. Subsequently, read group information was added, duplicates were flagged and the alignments were indexed. Short intronic overhangs were hard clipped using the SplitNCigarReads function and base quality score re-calibration was performed. Variants were called using the Haplotype Caller while ignoring soft-clipped bases and requiring a minimum phred-scaled confidence threshold of 20 as recommended in the Best Practices workflow. Clusters of at least three SNV calls in a 35 bp window were filtered. Furthermore, variants with Fisher Strand values greater than 30, which indicates significant strand bias, and Qual By Depth Values less than 2, which is a re-scaled quality score given the read depth, were filtered as well.

Germline variants called in the donor-matched bulk sequencing file and common polymorphisms present in dbSNP build 138 or the 1000 Genome Phase 1 panel were flagged and not considered during analyses relating to somatic mutations and processes. However, these variants were used to compute the relative fraction of RNA-confirmed germline SNPs to obtain a sensitivity estimate for RNA-confirmation of known genomic variants.

Mutation calls for the ERCC controls was performed in all available RNA-seq with a minimum of 10,000 reads aligned to the ERCC sequences. The workflow was identical to the SNV calling for endogenous RNA-seq data with the exception of germline and polymorphic variant flagging, which was not possible for the artificial ERCC control sequences²⁴⁹. SNV calls on ERCC sequences that passed GATK variant filtering were used to visualise the mutational spectrum of technical artefacts in RNA-seq data.

Germline SNP Calling in Bulk WGS Data

Germline SNPs from bulk WGS data were called using the GATK-v3.4. Haplotype Caller in single sample mode following the corresponding Best Practices Workflow. Read groups were added to the bwa genomic alignments with removed duplicates. After recalibration of base quality scores, gvcf files were produced using Haplotype Caller and genotype likelihoods computed in individual runs per donor. For the variant score calibration, dbSNP build 138 was specified as true known variants and the polymorphic sites from the 1000 Genome Phase 1 panel, HapMap 3 and Omni2.5 SNP array were used as training set. The flags considered for variant calibration were as recommended in the Best Practices and included QualByDepth, FisherStrand, StrandOddsRatio, MQRankSum and ReadPosRankSum. To maximise detection of germline SNPs, the desired sensitivity was set to 99.5% as defined by the GATK tranches. All variants with calibrated scores that passed the sensitivity-tranche filtering were used as donor-specific germline SNPs in further analyses.

Somatic SNV Calling in Single Cell WGS Data

Four SNV calling algorithms with different statistical framework and corresponding filtering routines were used for analysis of the WGS data. The four tools considered were Caveman, joint-sample Haplotype Caller from GATK, SCCaller and LiRA^{104,206,258,259}. All algorithms were run according to the corresponding guidelines and recommended thresholds for statistical significance were applied if appropriate. Following significance and any additional recommended filtering, all germline SNPs from the donor-matched bulk sequencing data and variants within the segmental duplication or simple repeat track from the UCSC Genome Browser database were removed as well²⁶⁰. For some of the subsequent analyses, a consensus set of base substitutions per cell was defined by all SNV calls that passed the appropriate filters of at least two different algorithms. The main principles and applied filters for the four different algorithms are briefly summarised in the following paragraphs.

Caveman is an expectation maximisation-based algorithm that uses a Bayesian classifier to decide for the most likely genotype at any given locus²⁵⁹. Caveman considers the ploidy and purity of the sample and to account for the non-uniformity of MDA-based amplification, the ploidy was set to five and purity to 90%. Caveman was run within the CASM Core Informatics Processing Pipeline at the Wellcome Sanger

Institute against the donor-matched bulk sequencing sample. Several post-processing filters are applied within this pipeline addressing quality metrics such as depth, base quality, read position and orientation (as described in ¹⁸). Germline variants are removed from Caveman calls during the pipeline processing through comparison to the matched bulk sequencing data and an unrelated normal panel to account for common polymorphisms. Further post-hoc filters were applied to remove variants if the supporting reads had median alignments scores (ASMD) less than 140 or more than half of them were clipped during alignment.

SNV calling using GATK was performed similar to the procedure described for bulk sequencing data. The main difference was the common genotyping of all ten single cell samples per donor to leverage any potential support of clonal variants across multiple cells. After variant score calibration, the 95% sensitivity tranche was selected to reduce the potential for false positives compared to the 99.5% tranche selected for germline variant calls ²⁵⁸.

SCCaller is a single cell-specific variant caller that was designed to detect locus-specific ADO and non-uniformity of amplification introduced by MDA ¹⁰⁴. SCCaller was used for variant calling in the MDA implementation chapter and is described in more detail in the corresponding method section.

LiRA is another single cell-specific variant caller that applies read-backed-phasing of potential somatic variants with germline heterozygous SNPs (gHet) from a donor-matched bulk reference. LiRA requires both observed alleles of somatic variants to be in perfect phasing agreement with the two alleles of the nearby gHet. The perfect phasing of both alleles is required to ensure that the somatic variant was present on the Watson as well as on the Crick strand of the original DNA molecule. Since strand information is lost during library preparation and sequencing, LiRA estimates a sample-specific composite coverage threshold that corresponds to the minimum read depth in bulk and single cell sequencing data at which somatic SNVs can be called. The main reasoning for this minimum read depth in single cells is that with increasing coverage the sole presence of phasing-concordant reads is not only due to undersampling of the MDA-amplified sequencing data. Despite the great non-uniformity of MDA amplification, LiRA assumes even binomial sampling from the

original Watson and Crick strand. Since LiRA can only call the fraction of somatic variants with a nearby gHet, the genome-wide burden is extrapolated on the basis of fraction of the genome with sufficient power and the corresponding amount of passed variant calls ²⁰⁶. LiRA was run according to the author's recommendations ²⁰⁶. Briefly, potential variants in single cell and gHets in bulk-sequencing data were called using the GATK Haplotype Caller ²⁵⁸. The raw variants before variant score calibration were used as input for LiRA. LiRA was run with references to the dbSNP build 151, the 1000 Genome Phase 1 panel and SHAPEIT v2.r837 as phasing software ¹⁶⁹. As recommended, 100 bootstrap iterations were specified during somatic rate estimation and the composite coverage threshold corresponding to a 10% FDR was estimated by LiRA. The actual variant calls from LiRA were considered for mutational spectrum analysis and the extrapolated genome-wide burden was compared to the total burden called by the three other variant callers.

Analysis of Shared Mutations from WGS Data and Phylogenies

Shared mutations were defined within the consensus SNV calls per cell. The amount of mutations shared between at least two or at least three cells per donor was considered. Additionally, the mutations shared across three cells were further restricted to sites with coverage of at least 10 reads and a maximum VAF of 0.25 in the donor-matched bulk sequencing data.

Mutations shared across at least two cells and the restricted set shared across at least three cells were subjected to phylogenetic reconstruction using the single cell-specific tree inference algorithm SCITE ²⁶¹. The average ADO rates per brain were used as false negative rates and the false positive rate was fixed at 10%. The maximum likelihood implementation to construct cell-centric trees was run in five independent Markov chains per donor and set of shared mutations with one million iterations each. All trees per run that achieved a similar maximum likelihood were emitted by SCITE and combined into consensus trees using the R package ape v5.3 ²⁶².

Integration of SNV Calls in RNA-seq and DNA-seq Data

SNV calls from RNA-seq data were integrated with the corresponding consensus SNV call set from WGS of the same single cell. Identical locus and alternative allele were required to be included in the putative set of integrated DNA-and-RNA calls. All

mutations within this putative set were visually checked in JBrowse to identify mutations supported by noisy reads, alignments errors or loci in close proximity to oligonucleotide repeats ²⁶³. Only mutations that could pass visual inspection were considered in the final set of integrated calls.

Estimation of Upper Limit for Genome-wide Burden of Mutations

An approximate estimate for the upper limit of the genome-wide burden of mutations was derived from the integrated DNA-and-RNA variant set. To estimate the fraction of the genome with sufficient power for SNV variant calls, the overlap of RNA-seq and WGS data from the cell was computed. The overlap was restricted to a coverage of at least eight reads in the WGS data since Caveman and SCCaller use read depth as hard filter. Given the fraction of the genome f_{DR} with sufficient power to call SNVs in an integrated manner and the corresponding amount of calls n_{DR} , an estimate for the genome-wide burden B was obtained. Furthermore, the RNA-confirmation rate of known genomic variants r_R (derived from the heterozygous germline SNP confirmation rate in RNA-seq) as well as a false negative rate for WGS calls r_D (derived from ADO rates computed on heterozygous germline SNPs in WGS data) were considered:

$$B = \frac{\frac{n_{DR}}{f_{DR}}}{r_R} \times \frac{r_D}{2}$$

Genome Feature Annotation

The fraction of the genome with sufficient coverage for integrated DNA-and-RNA variant calls as well as the corresponding integrated SNV calls were annotated with respect to their genome feature distribution using the R package *genomation* ²⁶⁴. The considered genomic features were promoter – defined as 1 kb window around transcription start sites – as well as exons, introns and intergenic regions as defined by the Ensembl version 73 annotation.

Association Between DNA-and-RNA SNV Calls and Gene Expression

The final set of integrated DNA-and-RNA variant calls were associated with a gene if they were located within the gene body as defined by the Ensembl version 73 annotation. To avoid inflated expression percentile ranks, only genes with FPKM values of at least one were used for the cell-specific ranking of expression values.

Mutational Spectra and Signature Extraction

The concept how mutational spectra were derived for base substitutions is described in the methods section of the MDA implementation chapter. Mutational signatures were extracted *de novo* from these spectra using SigProfiler v.2.1. that applies a non-negative matrix factorisation approach. If appropriate, the *de novo*-extracted signatures were compared to the SBS signatures from the ICGC PCAWG Platinum release using SigProfiler v.2.1 ³⁶. During this comparison, SigProfiler attempts to reconstruct the *de novo*-observed signatures through combinations of variable exposures to known signatures. To avoid overfitting the *de novo* signatures through fractional exposure to many known signatures, the stability of the initial solution is evaluated by comparison to bootstrap samples of the input mutational spectrum. An inverse relationship between quality of approximation and stability is usually observed and was evaluated for combinations of 1 – 10 known signatures. The solution with maximised approximation and stability was chosen and exposure to known signatures was considered significant if the combination of exposures and the *de novo*-extracted signatures displayed cosine similarity values greater than 0.95.

Gene Ontology and Pathway Enrichments

Gene Ontology (GO) and Pathway Enrichments were performed by GREAT v4.0.4 within the rGREAT package ^{265,266}. The GRCh37 annotation was used and backgrounds specified as appropriate and explained in the respective result sections. Since GREAT expects coordinate-based input, all genes considered for the GO analysis were converted into the transcription start site (TSS) coordinates as downloaded from the GREAT annotation. The coordinate-based input was mapped back to the corresponding gene within GREAT by choosing the association with the single-nearest gene within a 1 kb window while excluding curated enhancer regions. Considered GO categories were GO Biological Process, GO Cellular Component, GO Molecular Function as well as Panther, BioCyc and MSigDB Pathways.

Statistical Analysis

All statistical analysis was performed in R version 3.5.0 using core distribution functions unless otherwise indicated ¹⁷².

Results

Fifty Neurons are Selected for the Multiomics SNV Calling Approach

Single neuronal nuclei from the prefrontal cortex of five deceased neurotypical donors were FACS-sorted in three 96-well plates and subjected to G&T-seq using MDA for WGA of the nuclear DNA. Notably, the five donors were processed in three different batches with Brain19 and Brain20 constituting the first, Brain21 and Brain22 the second and Brain23 the third batch.

A previous publication focussed on SNV analyses in human neurons suggested that active transcription positively correlates with the mutational burden in post-mitotic neurons, which is contrary to SNV profiles observed in the human germline or cancer ^{103,267}. Since they used conventional MDA approaches, the authors from this previous study had to rely on average neuronal expression values from databases ¹⁰³. As neurons were shown to be a cell type with highly diverse expression profiles, the RNA layer from G&T-seq was used to select different neuronal subtypes to directly correlate the impact of active transcription with mutational burden ²⁶⁸⁻²⁷¹. In order to identify the neuronal subtypes in the present G&T-seq data, a gene expression count matrix against the Ensembl version 73 annotation was generated from the RNA-seq alignments of all 288 neuronal nuclei. The count matrix was restricted to 22,520 features with at least five counts in at least two nuclei. Furthermore, only nuclei with at least 100,000 total counts, a fraction of no more than 25% of these counts mapping to either mitochondrial genes or ERCC spike-in controls and a minimum of 1,000 expressed endogenous features were considered for the cell type identification. The final count matrix with 157 neuronal nuclei from all five donors was CPM-normalised and subjected to an unsupervised consensus clustering approach using the SC3 R package ²⁵⁷. Best consensus clustering results were obtained with three clusters with a corresponding average silhouette index of 0.83. The first cluster only contained three neuronal nuclei from Brain19 and Brain20 each, while the second and third cluster were comprised of 77 and 74 cells from all five brain donors, respectively (Figure 20A).

To associate known neuronal subtypes with the unbiased identification of subtypes, marker genes for these clusters were derived using the SC3 R package and compared to subtype-specific marker genes from the literature. When using the binary

classification approach for marker gene identification from SC3, ten genes each were identified as significant for the first two clusters and eight genes for the third cluster (AUROC > 0.8, Wilcoxon signed rank p-value < 0.1). Notably, the six neurons from the small first cluster also showed expression of all marker genes of the second cluster when restricted to the five most significant hits (Figure 20B). When these five most significant marker genes per cluster from the present G&T-seq data sets were compared to known marker genes from the literature, several overlaps were detected. One of the two major clusters expressed the glutamate transporter *SLC17A7* and *SV2B*, which is a synaptic vesicle protein preferentially found in excitatory neurons ^{269,270,272}. Marker genes for the other major cluster comprised the *GAD1* and *GAD2* genes, which are both coding for central enzymes in GABA synthesis, and the *GABA* transporter *SLC6A1* ²⁶⁹⁻²⁷¹. Therefore, these clusters could be identified as glutamatergic projection neurons and GABAergic interneurons, respectively. The five most significant marker genes for the small cluster of six neurons were not reported as neuronal marker genes in the literature. However, they expressed the five most significant marker genes of the major excitatory cluster as identified by SC3 and were found to cluster evenly within the population from the major excitatory cluster in a PCA based on their global expression profiles (Figure 20C). Therefore, the small cluster was considered to consist of a subtype of excitatory neurons. In summary, the resolution provided by the G&T-seq data set of 157 quality-filtered cortical neurons was sufficient to distinguish between excitatory and inhibitory neurons as main neuronal subtypes in the human cortex ^{232,233}.

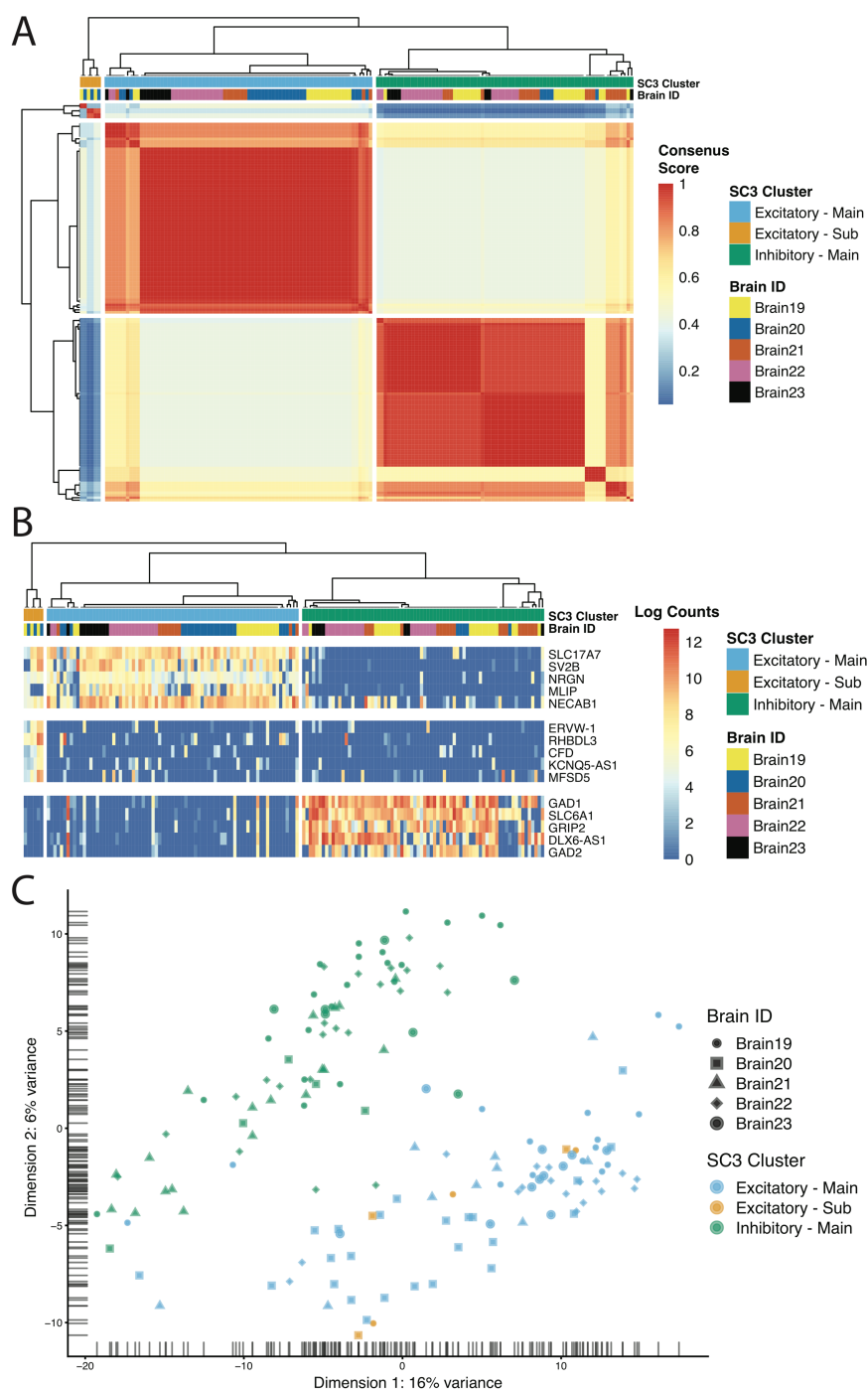


Figure 20: Unbiased Consensus Clustering Identifies Excitatory and Inhibitory Neurons as Main Subtypes in Transcriptome Data from G&T-seq. (A) Consensus clustering for cell type identification using SC3 reveals two major clusters with 77 and 74 neurons and an additional small cluster consisting of six neurons. **(B)** *De-novo* marker gene identification reveals *SLC17A7* and *SV2B* as highly expressed in one of the major as well as the small consensus cluster. Moreover, the marker genes discovered for the second major cluster comprise *GAD1*, *SLC6A1* and *GAD2*. These genes are ubiquitous marker genes for excitatory and inhibitory neurons. **(C)** The six neurons from the small cluster are evenly distributed within the main excitatory cluster on a PCA based on single nuclei RNA-seq data. Therefore, the small cluster is identified as a sub-cluster of excitatory neurons without previously published marker genes.

In parallel to the neuronal subtype identification, all MDA products were evaluated for their quality by comparing GbS-derived genotypes of MDA products to reference genotypes obtained from bulk sequencing of donor-matched brain samples. A total of 126 polymorphic sites across all chromosomes was compared for every MDA product and considerable difference was observed between processing batches. Neuronal nuclei from Brain19 and Brain20 featured an average of 45% (SD = 16%) concordance to the donor-matched bulk reference, while the Brain21 and Brain22 processing batch and nuclei from Brain 23 reached average concordance of 56% (SD = 17%) and 62% (SD = 15%), respectively. Notably, some nuclei from each batch also displayed drastically higher or lower concordance values (Figure 21).

Since GbS was demonstrated to be a powerful predictor of WGS quality, five neurons with the highest GbS quality score each from the main excitatory and main inhibitory neuron cluster were chosen per donor. MDA products from the fifty selected cells were subjected to amplification-free library preparation and WGS. The corresponding cDNA of selected MDA products was subjected to a conventional Illumina library preparation and further RNA-seq was performed in an attempt to improve read depth and transcriptomic coverage compared to the initial shallow RNA-seq for cell type identification.

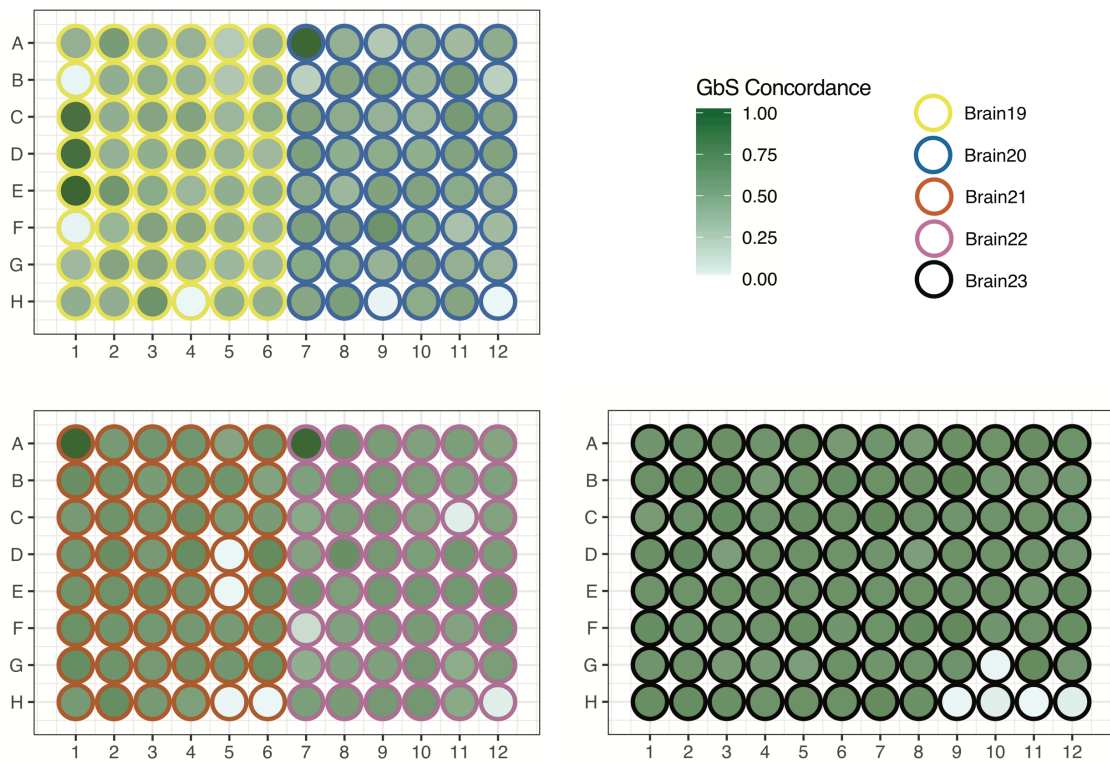


Figure 21: Quality Control of MDA Products Indicates WGA-related Batch Effects and Identifies the Most Suitable Amplification Products for WGS. MDA was performed in three separate batches as indicated in the plate layouts. Genotype concordance of MDA product was compared to germline references derived from donor-matched bulk sequencing data. While all processing batches include MDA products with low concordance values, the average GbS-indicated quality score is substantially lower for the processing batch including Brain19 and Brain20. The five excitatory and inhibitory neurons with the best GbS score per donor were selected for WGS to generate the final data set of 50 single cells considered in further analyses.

Increased Read Depth Slightly Improves Feature Detection in RNA Data

The initial RNA-seq for neuronal subtype identification had been performed on Nextera-based libraries on two Illumina HiSeq 2000 lanes for all 288 neuronal nuclei. To improve transcriptome-wide coverage for the fifty neurons selected for WGS, their corresponding cDNA was subjected to conventional Illumina sequencing library preparation and re-sequenced on another two Illumina HiSeq 2000 lanes. When the obtained read depths for the fifty selected neurons were compared between the two rounds of RNA-seq, an increase of the median depth from about $8,5 \times 10^5$ to nearly $1,1 \times 10^6$ reads per nuclei was observed. Moreover, the median number of expressed endogenous features per nuclei was slightly increased with about 5,800 features detected for the deeper sequencing depth compared to the initial round of RNA-seq.

with only 5500 features (Figure 22). Therefore, conventional RNA-seq library preparation increased read depth and improved transcriptomic coverage. The corresponding data was used in further analyses.

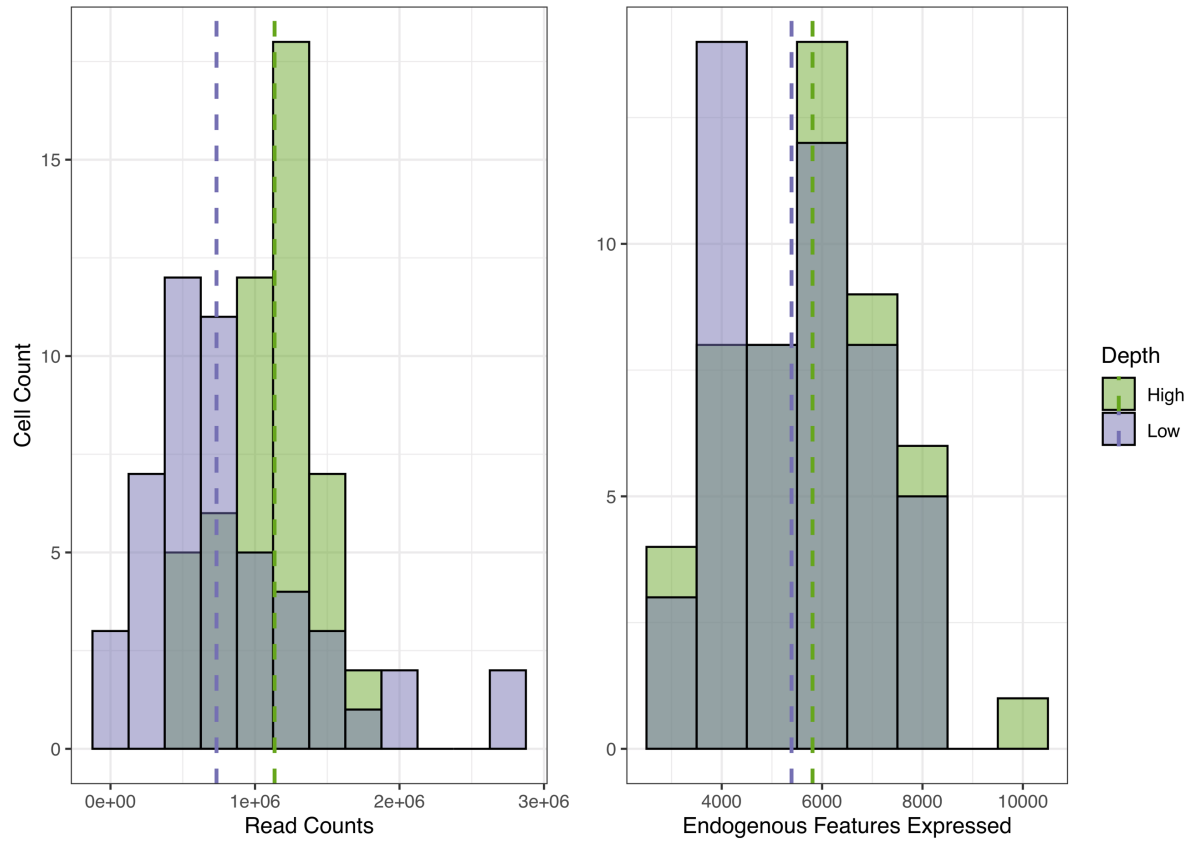


Figure 22: Deep Sequencing of Transcriptomes Corresponding to the 50 MDA Products Selected for WGS Improves Endogenous Feature Detection. New conventional Illumina sequencing libraries were generated from the cDNAs of 50 neurons that were selected for WGS based on the initial neuronal subtype identification and GbS quality scores. The deeper sequencing improved read depth by an average of over 250,000 reads per nuclei and resulted in an average detection of about 5,800 expressed features per nuclei compared to about 5,500 features for the shallow RNA-seq.

Deep-sequenced RNA Confirms Excitatory and Inhibitory Neuronal Subtypes

Since increased read depth for the transcriptomes of neurons selected for WGS improved the detection of expressed features per nucleus, the neuronal subtype classification was repeated using the combined data set of shallow and high sequencing depth RNA-seq. Similar to initial analysis, the combined data set was analysed using SC3 and the best intra-cluster cohesion and inter-cluster separation was achieved with three consensus clusters with a silhouette value of 0.83

(Figure 23A). The best consensus clustering identified two major clusters with 104 and 95 nuclei as well as a small cluster with nine nuclei. While the two major clusters contained neurons from all brains and sequencing depths, the small cluster included three high-depth transcriptomes from Brain19 and Brain20, their corresponding low-depth transcriptomes and three additional low-depth transcriptomes from Brain21 that had not been selected for the deep RNA-seq. Marker genes for the three consensus clusters were predicted by SC3 as described for the analysis of the shallow sequencing depth and restricted to the five most significant hits per consensus cluster (Figure 23B). Compared to the previous marker gene analysis, three of the current top five hits for the major cluster with 104 neurons overlapped with the cluster identified as excitatory projection neurons, including the two published marker genes *SLC17A7* and *SVB2*^{269,270,272}. This cluster contained the 25 deep transcriptomes of excitatory neurons for which the corresponding MDA products were selected for WGS. Four marker genes from the major cluster with 95 nuclei overlapped with the five most significant hits from the cluster previously identified as inhibitory interneurons. This included *GAD1* and *SLC6A1* as known marker genes for GABAergic neurons²⁶⁹⁻²⁷¹. Marker genes for the small cluster did not overlap with the previous marker gene analysis from the shallow sequencing and did not contain any published neuronal marker genes. The three deep transcriptomes within this small cluster all featured expression of *GAD1* and *SLC6A1* (Figure 23B). Moreover, all neurons from the small consensus cluster were found to cluster evenly within the major inhibitory cluster when visualised on a PCA (Figure 23C). Since the small cluster still displayed expression similar to GABAergic neurons but could not be identified as a further particular subtype, the 25 deep transcriptomes from the small and the major inhibitory consensus cluster were grouped again for further analyses.

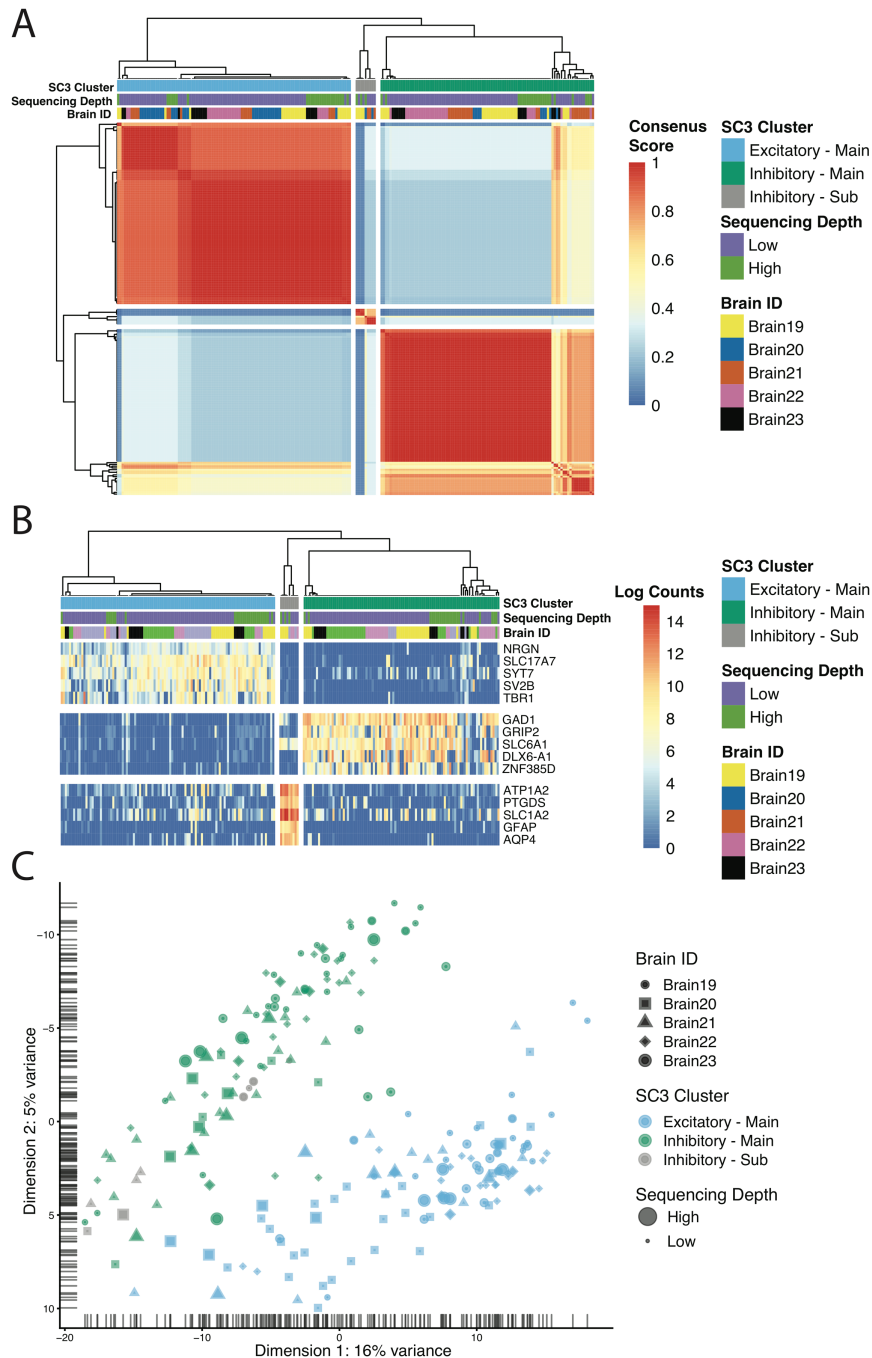


Figure 23: Excitatory and Inhibitory Neurons are Still Identified as Main Subtypes when Deep Sequenced Transcriptomes are Considered. The unbiased cell type detection was repeated using the combined shallow and deep RNA-seq data to evaluate if further known subtypes can be defined with improved feature detection in deep RNA-seq data. **(A)** Consensus clustering using SC3 identified two major clusters with as well as a new small cluster comprised of nine cells. **(B)** Marker genes for the two main clusters defined by SC3 overlapped with the known marker genes for excitatory neurons *SLC17A7* and *SV2B* as well as with *GAD1* and *SLC6A1* for inhibitory neurons. The three deep sequenced cells from the small cluster also featured *GAD1* and *SLC6A1* expression but the unique marker genes for the small cluster did not overlap with any neuron-specific marker genes from the literature. **(C)** All cells from the small cluster are evenly distributed within the main excitatory cluster and thus, were grouped together for further analyses.

Germline SNP Confirmation Rate in RNA is Limited by Bimodal VAF Distribution

SNVs in the RNA-seq data were called and filtered using GATK while following the corresponding Best Practices workflow as described in the methods section ²⁵⁸. SNV calls in RNA-seq can derive from germline and somatic mutations of the genome, post-transcriptional modifications of the mRNA or represent technical artefacts. SNV calls from RNA-seq data were planned to be integrated with the corresponding calls from the genomic layer of the same cell in the presented G&T-seq SNV calling approach. Any calls representing post-transcriptional modifications of the mRNA or technical artefacts would be filtered during the integration due to their absence in the genomic call set. Therefore, high sensitivity for RNA-derived SNV calls was desirable to maximise the confirmation of genomic variants. An estimate for this sensitivity was calculated using the donor-matched germline SNPs known from bulk DNA sequencing data and common polymorphisms from dbSNP build 138 or the 1000 Genome Phase 1 panel. The amount of germline SNP loci that demonstrated any RNA-seq coverage was highly variable with averages of 20135 (SD = 6681), 9364 (SD = 7722), 8397 (SD = 5097), 13938 (SD = 6716) and 11942 (SD = 5737) for Brain19 to Brain23, respectively. When restricted to germline SNP loci covered by RNA-seq, a relatively constant fraction of 58% (SD = 6%) of these loci featured an RNA-seq SNV call with matching alternative allele (Figure 24A). However, an average of 91% (SD = 1%) of the variant calls were homozygous in the RNA-seq data with very little variation between cells or donors. In contrast, an average of only 40% (SD = 1%) of these germline SNPs were found to be homozygous in the donor-matched bulk sequencing data. When the VAF distribution in RNA-seq reads across all loci was visualised, a clearly bimodal pattern was observed with the majority of loci only displaying either the reference or an alternative allele (Figure 24B). Due to this intrinsic restriction of sensitivity, no distinction between heterozygous and homozygous SNV calls was made in further analyses. Furthermore, the average sensitivity of 58% to detect known genomic variants in RNA-seq covered loci was accepted to be the upper limit for this G&T-seq data set and no further maximisation was attempted.

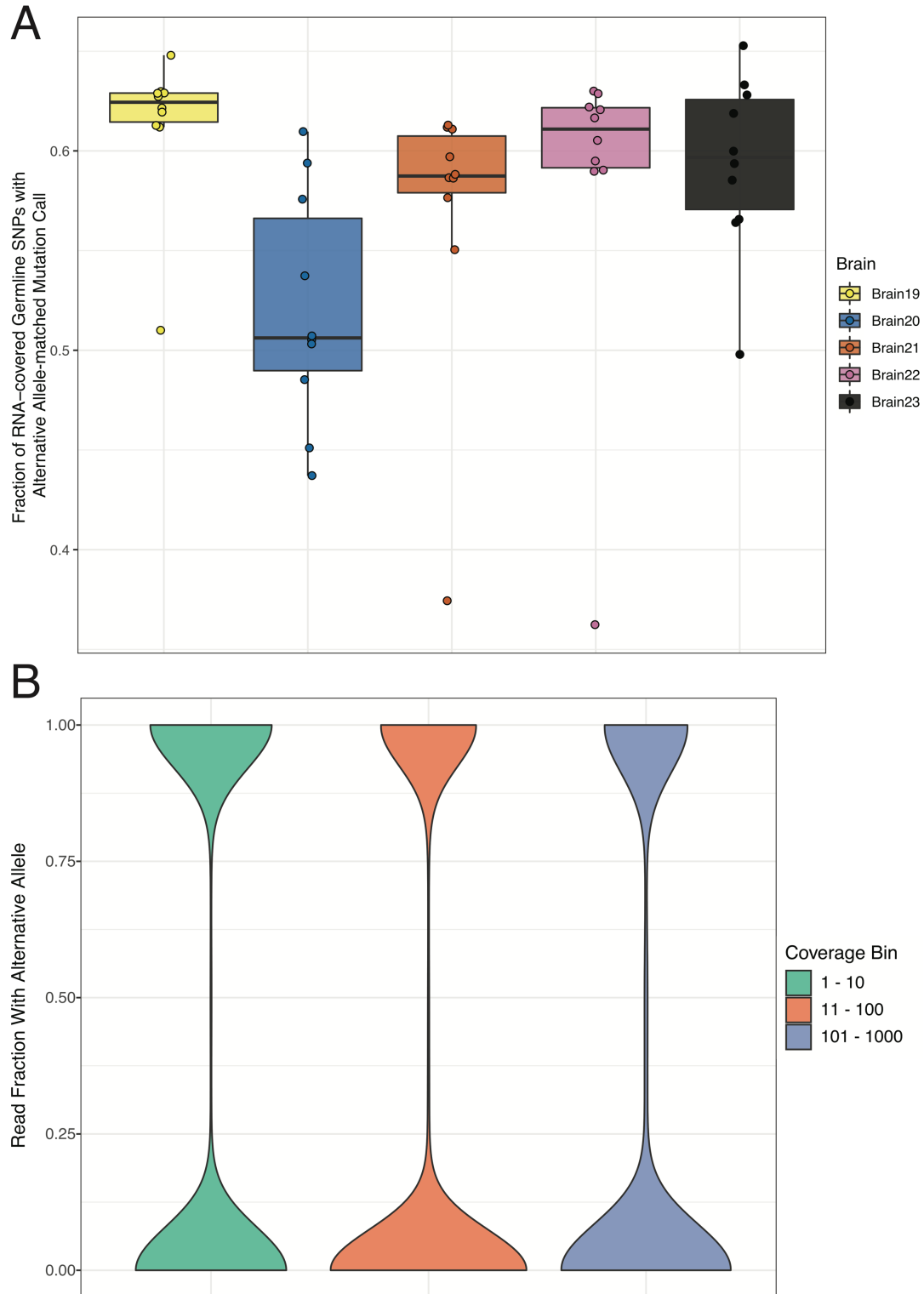


Figure 24: Germline SNP Confirmation Rate in RNA-seq Data is Limited by the Bimodal Pattern of Allelic Expression. (A) The detection rate of the alternative allele of germline SNPs derived from donor-matched DNA bulk sequencing was evaluated in SNV calls from single cell RNA-seq data. Notably, only SNP loci with RNA coverage were considered. While cell-specific variation exists, the maximum confirmation rate is limited to about 65%. (B) A clearly bimodal VAF distribution in single cell RNA-seq data was observed for all loci across all read depths, which intrinsically limited the germline SNP confirmation rate.

Substantial Level of RNA Editing Detected in Single Cell RNA-seq Data

Following the removal of germline variants and flagging of calls on the ERCC control sequences, substantial levels of base substitutions were detected with considerable intra- and interindividual variation. On average, 10,196 (SD = 3262), 3890 (SD = 4092), 4025 (SD = 2703), 7612 (SD = 3983) and 6363 (SD = 3271) SNV calls passed the filtering for Brain19 to Brain23, respectively (Figure 25A). Interindividual variation was found to be statistically significant ($F[4,45]$, $p < 0.001$, one-way Anova) and corresponding post-hoc tests demonstrated that the difference between Brain19 with Brain20 and Brain21 was below significance thresholds of 0.05 (adjusted $p < 0.003$, Tukey's HSD). Notably, the interindividual differences were still significant when correcting the total number of SNV calls in the RNA-seq data for sequencing depth and total genomic breadth of coverage. Moreover, SNV calls were found to be largely independent of sequencing depth, which suggested that no direct correlation exists between absolute expression values and the ability to call SNVs in the RNA-seq data (Figure 25B).

A mutational spectrum was generated for SNV calls on endogenous features for the fifty deep-sequenced transcriptomes and compared to the corresponding spectrum of base substitutions called on the ERCC control sequences (Figure 26). The mutational spectra displayed a modest correlation of 0.47 but greatly differed in the relative amount of the six main substitutions classes. Nearly 47% of SNV calls on endogenous features were classed as T>C with only about 23% for calls on ERCC control sequences (Bonferroni-adjusted $p < 1e-14$, Binomial test). The main form of canonical RNA editing in humans is adenosine deamination to inosine (A-to-I) mediated by the ADAR family²⁷³. Any inosine present on the original mRNA is detected as guanosine during Illumina sequencing and due to the classification based on the pyrimidine base of the Watson-Crick-base pair, any A>G mutation caused is classified as T>C mutation in the mutational spectrum. Due to the significant enrichment of SNV calls on endogenous features over the artefact background that correspond to the main canonical form of RNA editing in humans, a substantial level of RNA editing in the cortical neurons was implied.

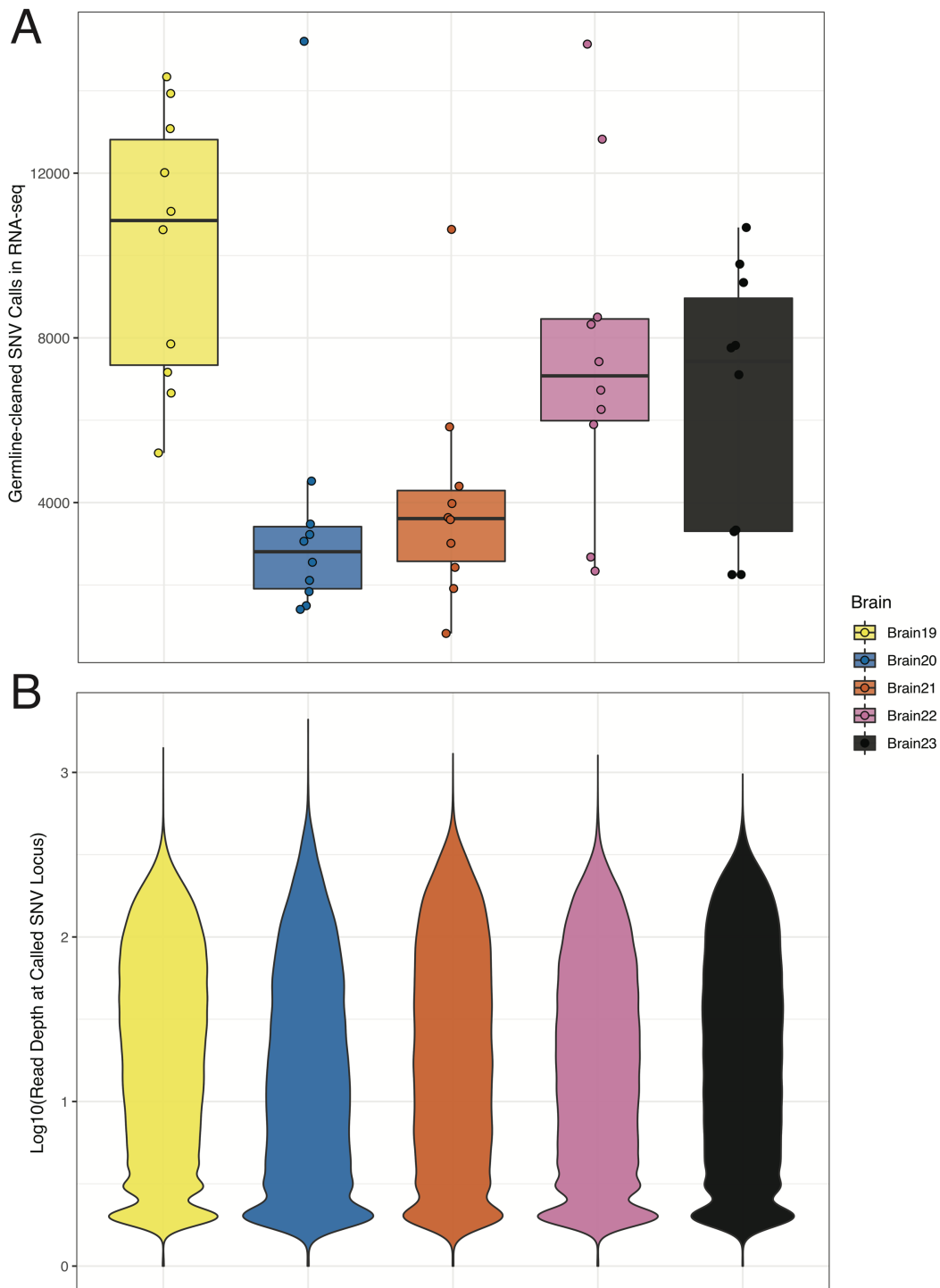


Figure 25: Substantial Amounts of Base Substitutions are Observed in Germline-cleaned SNV Calls from RNA-seq Data Across all Read Depths. (A) SNVs in single nucleus RNA-seq data were called following the GATK Best Practices and germline variants derived from donor-matched DNA bulk sequencing were removed. Substantial amounts of base substitution calls were observed, which account for the expression of somatic genomic variants, post-transcriptional modification of mRNA and technical noise. **(B)** SNV calls in RNA-seq data display a relatively even distribution across all read depths, which suggested that transcriptomic coverage did not limit SNV calling in single nucleus RNA-seq data.

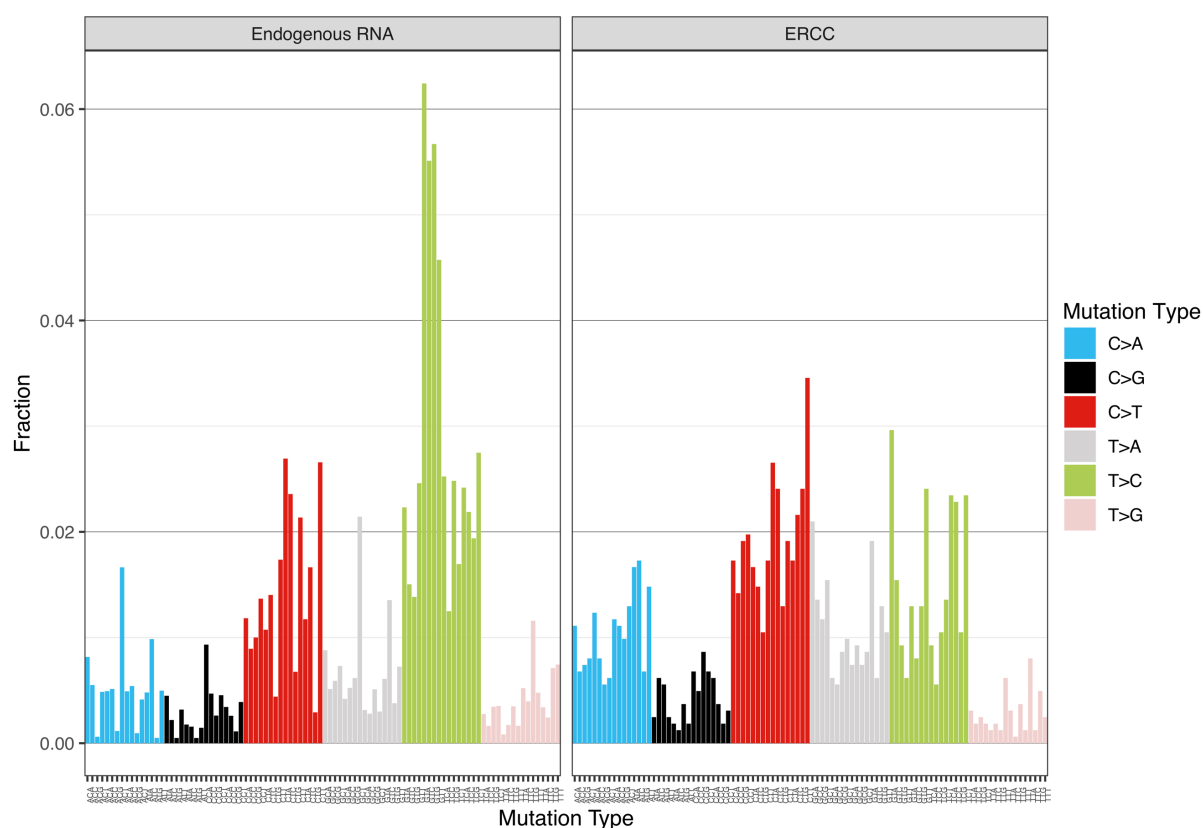


Figure 26: The Mutational Spectrum of Base Substitutions in Endogenous RNA is Significantly Enriched in T>C Substitutions Compared to Artefactual Calls on ERCC Control Sequences. Base substitutions were classified according to the pyrimidine base change of the Watson-Crick base pair and further distinguished by their immediate 3'- and 5'-context. The profile derived from SNV calls on endogenous RNA substantially differs from the profile generated by technical noise as estimated on artificial ERCC spike-ins. The most prominent difference is the enrichment of T>C substitutions in endogenous RNA calls (adj. $p < 1e-14$, Binomial test).

RNA Editing is More Prevalent in Excitatory than in Inhibitory Neurons

Since a substantial amount of RNA editing was indicated in the deep transcriptome data from cortical neurons, potential differences between the two main subtypes were considered. As the main enrichment over the artefact background was observed for T>C substitutions, the following analyses were restricted to SNV calls that were compliant with canonical A-to-I editing. SNV calls corresponding to A-to-I editing were significantly higher in excitatory neurons compared to inhibitory neurons with average counts of 3767 and 2242, respectively ($p = 4e-3$, Student's t-test, Figure 27A). Notably, the significant difference was maintained when controlling for read depth and breadth of genome coverage. Furthermore, a significantly higher fraction of genes expressed in excitatory neurons contained at least one T>C substitution within their

gene body compared to genes expressed in inhibitory neurons (odds ratio = 1.13, $p = 2.2 \times 10^{-6}$, Fisher's Exact test). To explore the functional impact of RNA editing in the neuronal subtypes, potential editing events were restricted to recurrent T>C substitutions that were called in at least two cells per subtype. This resulted in 3593 recurrent sites for excitatory and 1374 recurrent sites for inhibitory neurons. These recurrent sites were found within the gene bodies of 1834 genes for excitatory neurons and 737 genes for inhibitory neurons according to the Ensembl version 73 annotation. The genes associated with recurrent potential A-to-I editing sites were analysed for GO term and pathway enrichment compared to the background of all expressed genes in the corresponding neuronal subtype. No pathway enrichments could be detected but several GO Molecular Function terms were associated with the potentially edited genes for both neuronal subtypes (Figure 27B). When restricting to the five most significant hits according to the hypergeometric p-values, the GO terms for inhibitory neurons could all be related to calcium ion channel and transporter activity. While similar GO terms were also within the five most significant hits for excitatory neurons, two GO terms related to glutamate receptor activity, which is the main neurotransmitter of excitatory neurons. Therefore, A-to-I editing was indicated to affect common functions in both neuronal subtypes with additional subtype-specific molecular functions in excitatory neurons.

Canonical A-to-I editing is thought to be mainly mediated by enzymes of the ADAR family, which comprises the catalytically active ADAR and ADARB1 genes as well as the potentially inactive ADARB2 gene ²⁷³. Differences in expression of ADAR family members were considered for the observed differences in A-to-I editing between neuronal subtypes. However, when A-to-I editing compliant SNV calls per cell were compared to ADAR and ADARB1 expression, no significant effect could be detected in an ordinary linear regression (Adj. $R^2 = -0.03$, $p = 0.7$ for ADAR, see Figure 28; Adj. $R^2 = -0.07$, $p = 0.8$ for ADARB1).

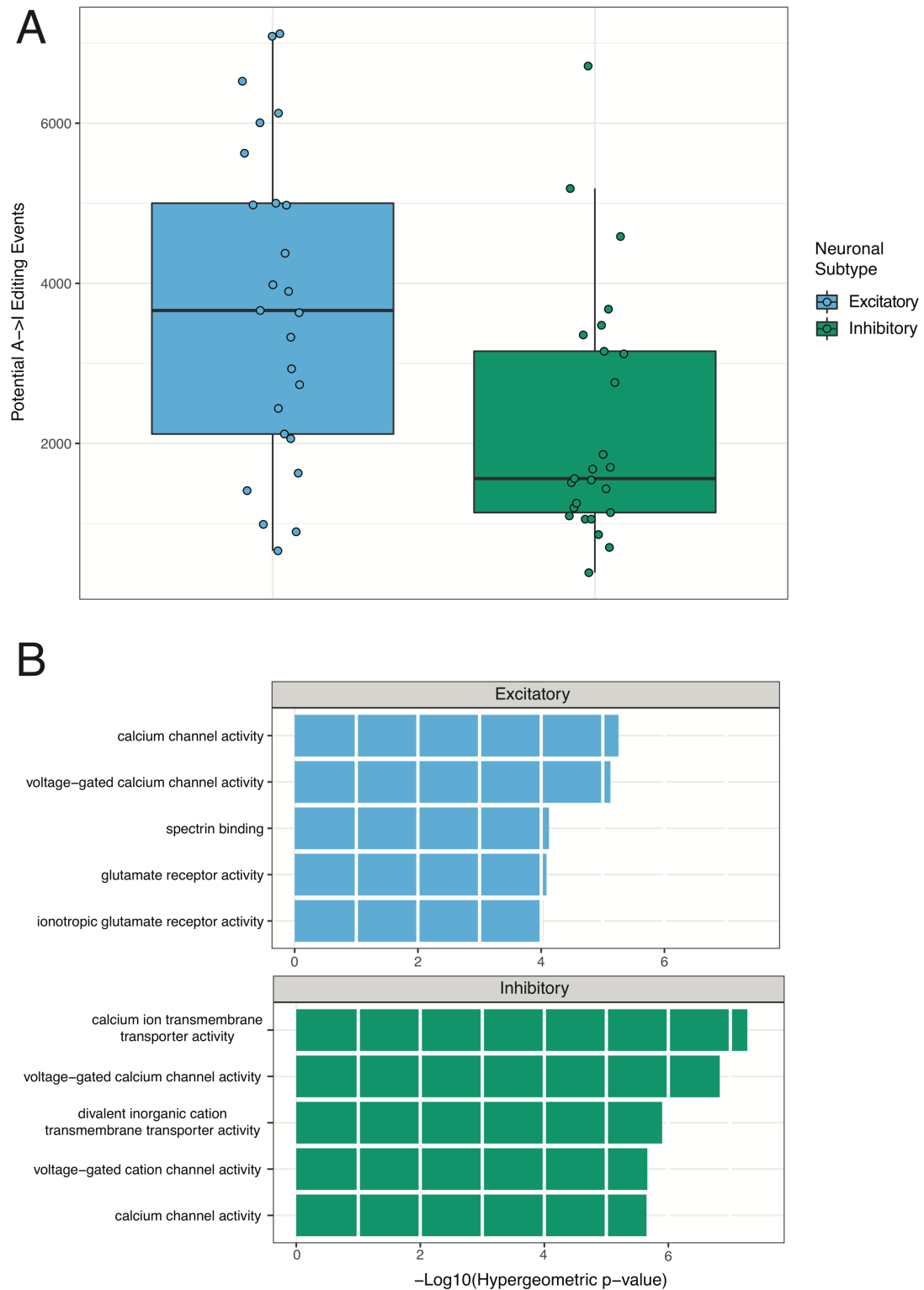


Figure 27: RNA Editing is More Prevalent in Excitatory Neurons and is Enriched in Genes Regulating Glutamate Receptor and Ion Channel Activity. (A) Substitution calls that were in accordance with canonical A-to-I RNA editing are significantly enriched in excitatory compared to inhibitory neurons ($p = 4e-3$, Student's t-test) (B) Genes containing loci with recurrent SNV calls in accordance with A-to-I editing were subjected to GO enrichment analysis. These genes were enriched in several Molecular Function annotations regarding ion channel activity for both subtypes as well as glutamate receptor activity for excitatory neurons.

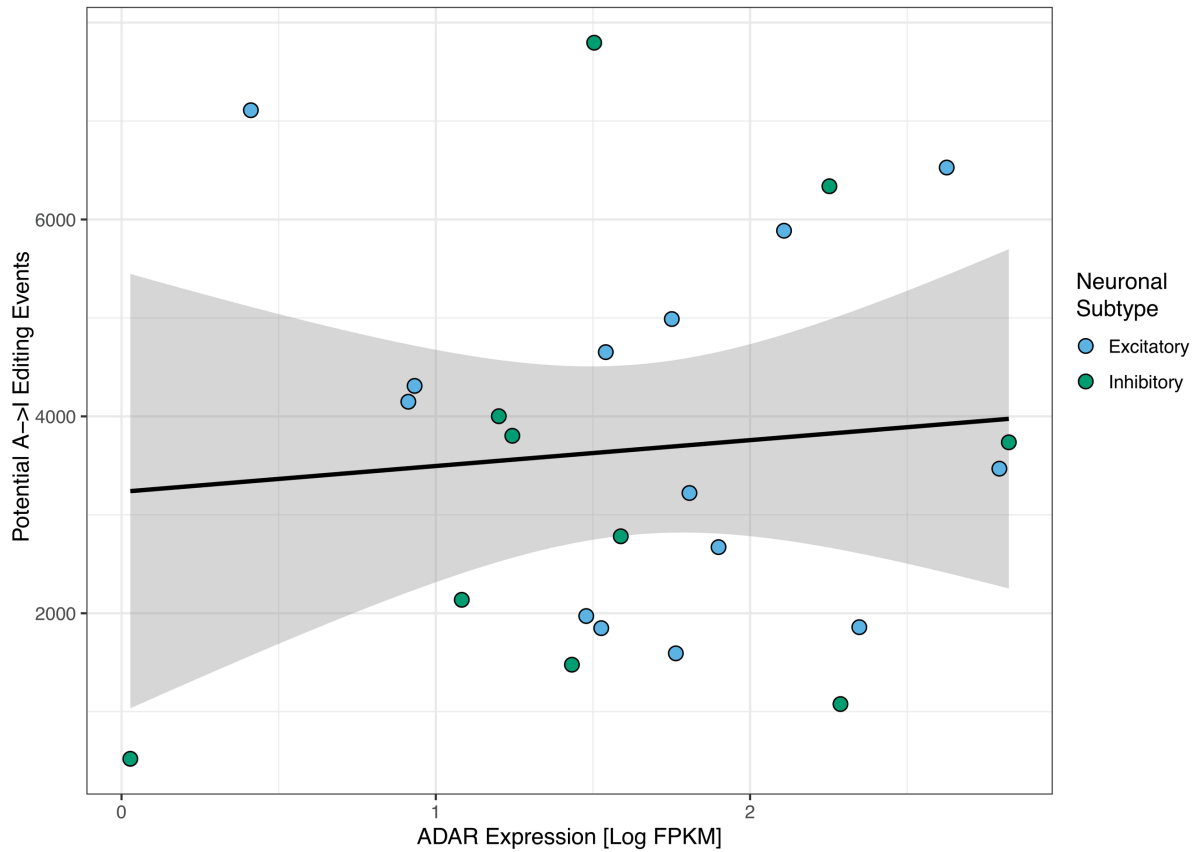


Figure 28: Differences in ADAR Expression Levels do not Account for Observed Variation in SNV Calls in Accordance with A-to-I Editing. An ordinary linear regression was performed to correlate differences in ADAR expression to suggested variation in A-to-I editing. However, no significant effect could be observed (Adj. $R^2 = -0.03$, $p = 0.7$).

Batch Effects and Extreme Mutational Burden are Discovered in WGS Data

The WGS libraries from the fifty selected MDA products were sequenced to an average depth of 33X (SD = 3X). While the mean coverage was similar across brains, the fraction displaying moderate to high read depth more suitable for SNV calling, varied significantly among brains (Genome Fraction covered at $\geq 15X$; $F[4,45]$, $p < 1e-9$, one-way Anova). On average, only 40% (SD = 5%) and 44% (SD = 5%) of the genome displayed read depths greater than 15X for Brain19 and Brain20, while a significantly higher proportion of 51% (SD = 3%), 54% (SD = 2%) and 52% (SD = 3%) of the genome featured similar coverage values for Brain21 to Brain23 (Figure 29A; adjusted $p < 0.002$ for all comparisons between Brain19 and Brain20 with Brain21, Brain22 and Brain23, Tukey's HSD). Moreover, MDA products from Brain19 and Brain20 displayed greater variance in ADO rates - as determined through comparison to germline SNPs derived from donor-matched bulk sequencing data - compared to Brain21 to Brain23 (Figure 29B). However, the nominally elevated average of 54% ADO (SD = 4%) for Brain19 and Brain20 was not significantly different to the average of 53% ADO (SD = 2%) for Brain21 to Brain23 ($F[4,45]$, $p = 0.7$, one-way Anova). Notably, MDA amplifications for Brain19 and Brain20 were performed in one processing batch and lower WGS quality was indicated by the lower GbS quality scores as described above (Figure 21).

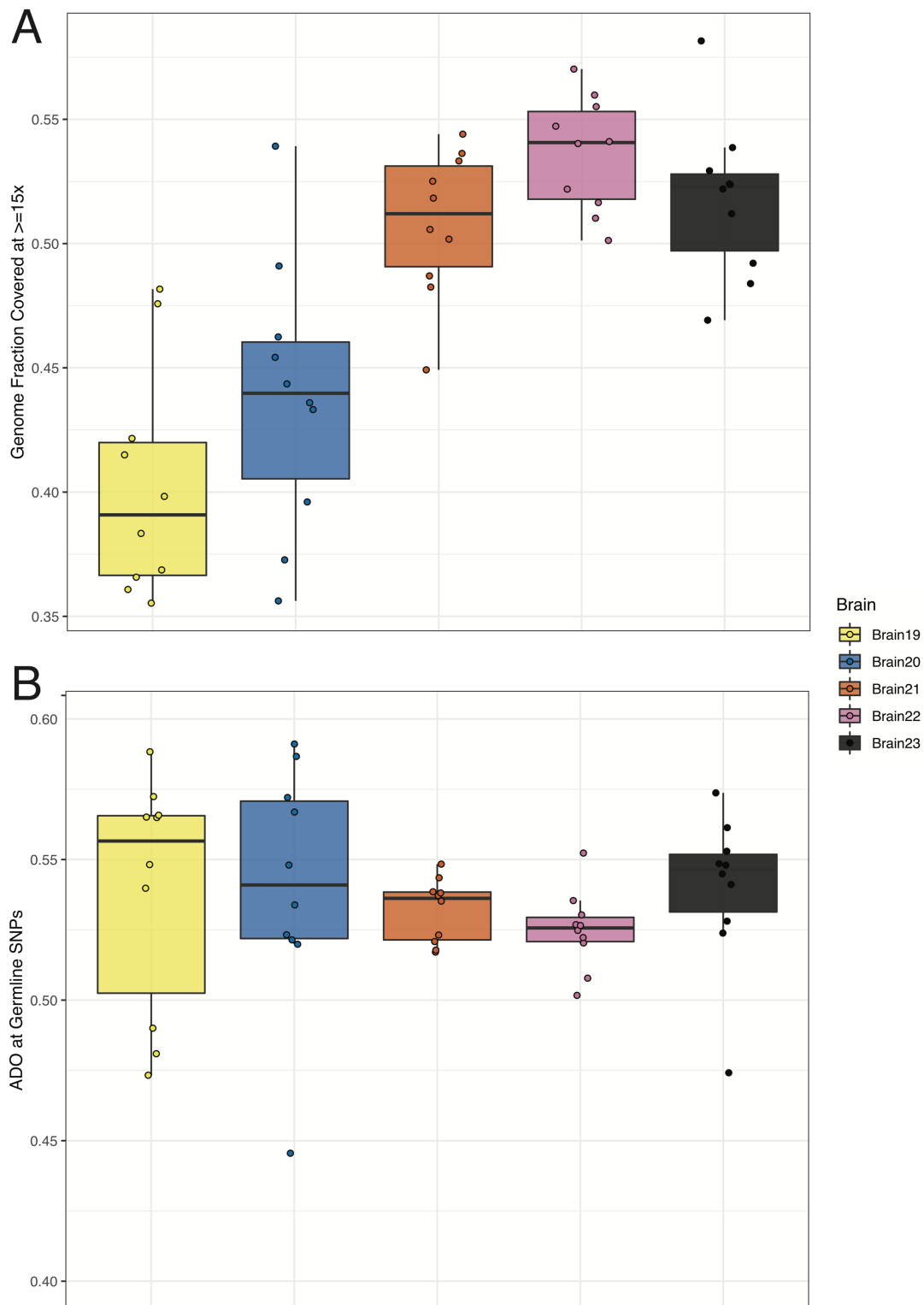


Figure 29: WGS from Brain19 and Brain20 Display Limited Breadth of Coverage and Substantial ADO Rates are Observed for all Selected MDA Products. (A) All MDA products were sequenced to the same average depth of 33X. However, Brain19 and Brain20 displayed significantly lower breadth of coverage at modest read depth compared to the remaining three donors. **(B)** ADO rates were evaluated on WGS covered loci of heterozygous SNPs known from donor-matched bulk sequencing. Substantial ADO rates above 50% for most cells were observed with greater variation for Brain19 and Brain20 compared to cells from the other three donors.

Despite the indicated differences in WGS quality, base substitutions were determined in WGS data of all MDA products using the four algorithms Caveman, GATK Haplotype Caller, SCCaller and LiRA as described in the methods section ^{104,206,258,259}. Subsequent to germline variant removal and further algorithm-specific filtering as described in the methods section, differences across an order of magnitude with respect to the total mutational burden were observed between different calling algorithms and between donors (Figure 30). Independent of the algorithm used, Brain19 and Brain20 featured substantially higher mutational burdens than neurons from the remaining three donors. For Caveman, the average mutational burden for cells from Brain19 to Brain23 were 361,898 (SD = 55,510), 277,521 (SD = 38,186), 73,950 (SD = 5,547), 65,978 (SD = 6,839) and 76,843 (SD = 8,562). Similar mutational burdens were estimated for Brain19 to Brain23 by GATK with 449,432 (SD = 60,679), 316,887 (SD = 37,325), 61,797 (SD = 3,652), 65,477 (SD = 5,909) and 75,998 (SD = 8,428) average mutations per cell of the corresponding donor. The single cell-specific algorithm SCCaller yielded overall drastically reduced numbers of mutations, while the general pattern between different donors was maintained. The average mutational burdens of 122,441 (SD = 41,053), 104,181 (SD = 21,705), 20,222 (SD = 1,705), 17,995 (SD = 1,413) and 20,381 (SD = 2,110) were the most conservative estimates obtained from all considered algorithms. The read-backed-phasing algorithm LiRA that was used to explore base substitution variation in MDA-amplified WGS data from cortical neurons in a previous study ¹⁰², indicated even higher mutational burdens for Brain19 and Brain20 than the other three callers, while estimates for Brain21 to Brain23 were lower than the burden called by Caveman or GATK but higher than for SCCaller. On average, LiRA estimated an average mutational burden of 610,448 (SD = 89,916), 443,599 (SD = 63,293), 49,790 (SD = 6,091), 42,448 (SD = 5,029) and 47,946 (SD = 10,407) for Brain19 to Brain23, respectively (Figure 30). Notably, the mutational burden reported in other normal human tissues including muscle, liver, colon, intestine, blood, dermal fibroblasts, endometrium as well as neurons, range between a few hundred up to several thousand base substitutions per cell ^{77,84,86,87,91,92,102-104,274}. The most extreme estimates for genome-wide mutational burden by LiRA for neurotypical cells from Brain19 with an average of over 600,000 base substitutions are even high compared to most cancer samples ²⁶⁷.

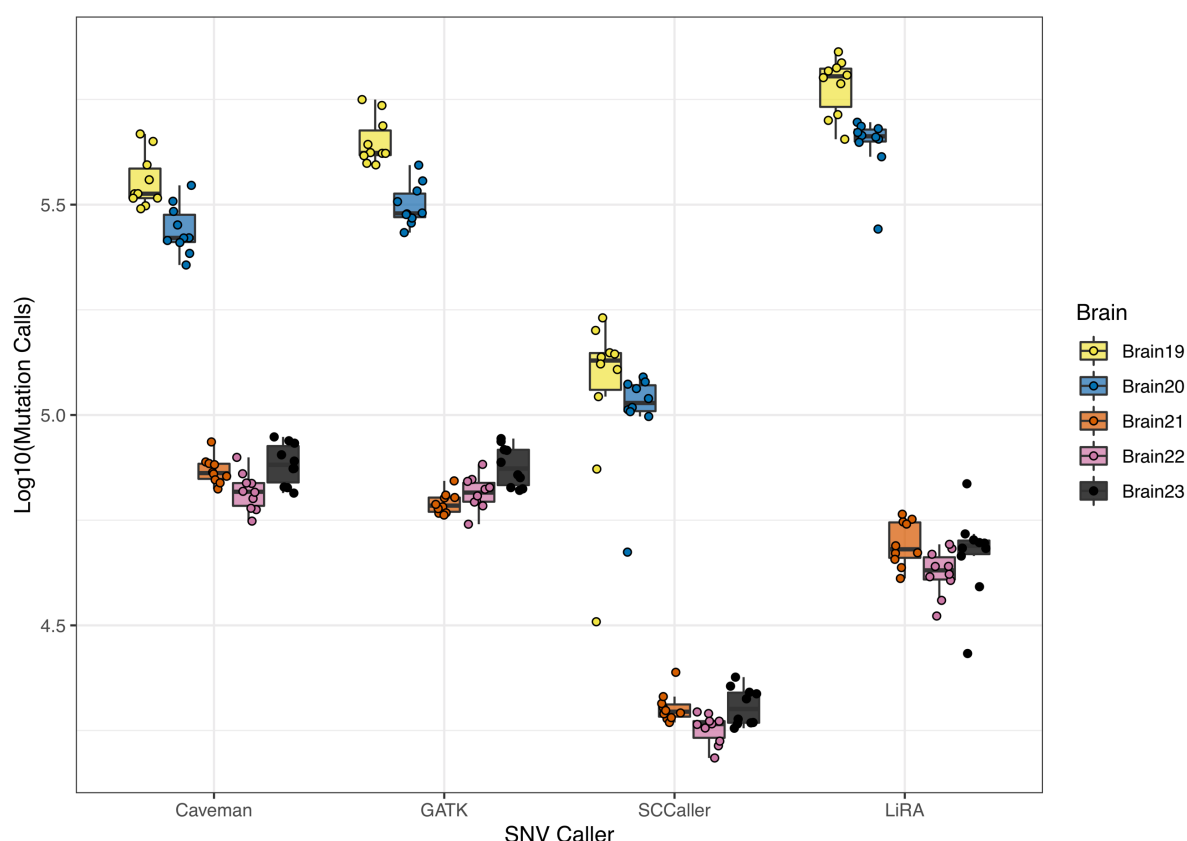


Figure 30: Extreme Mutational Burdens were Implied in Single Cell WGS Data. SNVs from WGS data were called using the four different algorithms Caveman, GATK, SCCaller and LiRA. While the results differed substantially between different algorithms, the mutational burden in Brain19 and Brain20 was implied to be nearly an order of magnitude greater than for the other three donors. Notably, Brain19 and Brain20 were processed in the same WGA batch.

Mutational spectra were generated for the genome-wide calls from the four different algorithms (Figure 31). The relative contribution of the 96 considered context-specific substitution classes displayed great similarity between the different algorithms and strong dominance of C>T substitutions. A total of 83%, 84%, 90% and 95% of the mutations called by Caveman, GATK, SCCaller and LiRA, respectively, were classified as C>T substitutions. Notably, the majority of cells from Brain19 and Brain20 displayed even greater proportions of C>T substitutions than the average across all brains. The majority of technical artefacts introduced during MDA-based WGA is known to be cytosine deamination that results into C>T variant calls ^{104,275}.

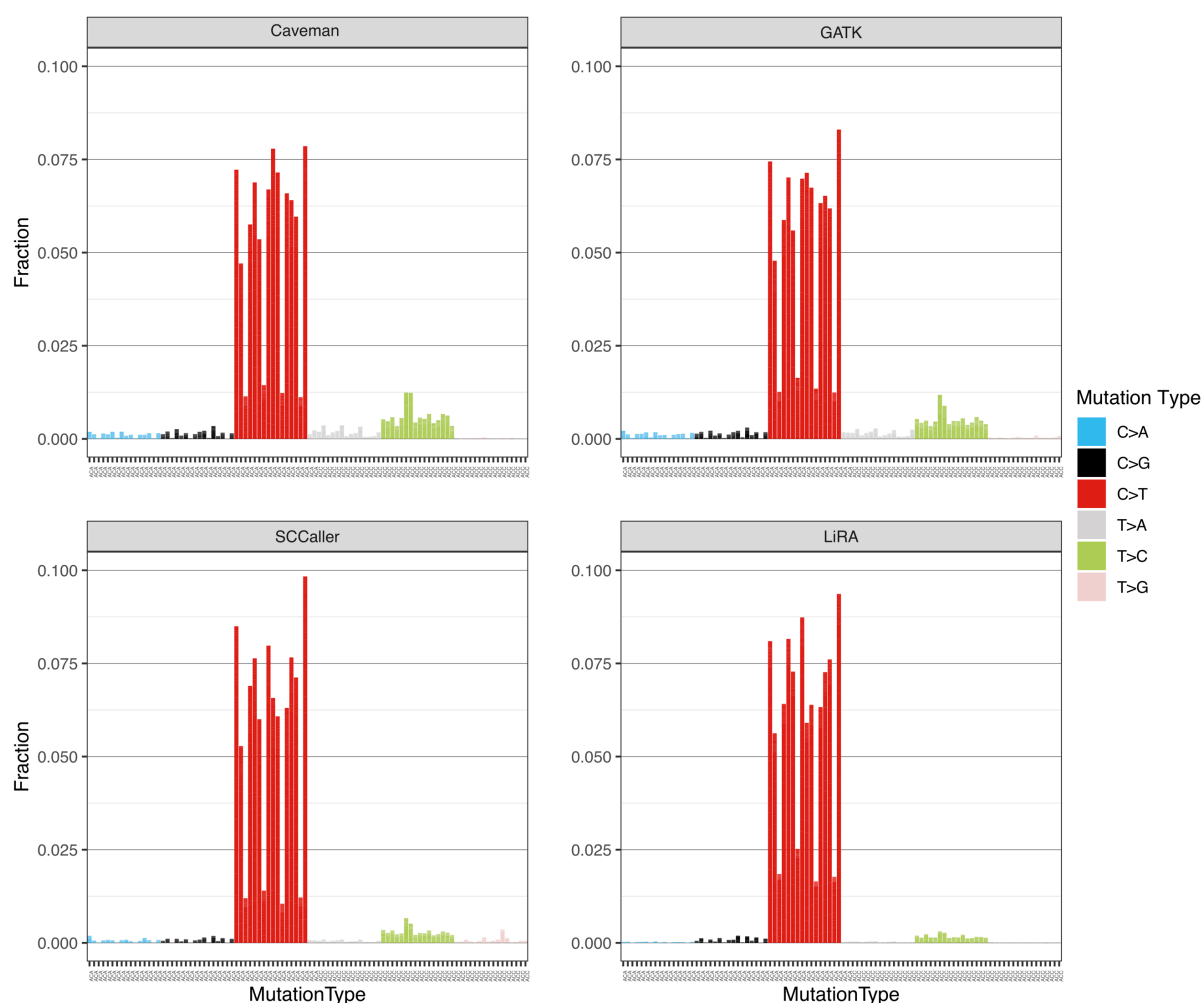


Figure 31: Mutational Spectra for SNVs Called from All Considered Algorithms are Strongly Enriched in C>T Substitutions. Base substitutions were classified according to the pyrimidine base change of the Watson-Crick base pair and further distinguished by their immediate 3'- and 5'-context. The profiles generated by the four different algorithms are highly correlated and all display overwhelming dominance of C>T substitutions.

The extreme mutational burden with strong enrichment in substitutions corresponding to the most prevalent WGA-associated artefact, implied technical artefacts as the primary source of SNV calls independent of the algorithm used. Additionally, it was reasonable to assume a WGA batch effect resulting in elevated levels of technical noise in Brain19 and Brain20 compared to the remaining three donors due to their even higher estimates of mutational burden and stronger enrichment in C>T variant calls.

Deconvolution of Mutational Spectra and Restriction to Shared Mutations Cannot Resolve the High Artefact Background in WGS Data

The mutational spectra from the four different algorithms were deconvoluted to extract underlying mutational signatures in an attempt to distinguish true biological processes from signature contributions driven by technical artefacts. Deconvolution of the mutational spectra into individual mutational signatures suggested the existence of SBS19, SBS30, SBS32 and SBS44 as defined in the ICGC PCAWG Platinum release for the mutational spectra from Caveman, GATK and SCCaller and only SBS19 and SBS30 for the corresponding LiRA spectrum ³⁶. SBS19 and SBS30 also accounted for the majority of the mutations when the additional existence of SBS32 and SBS44 was suggested, with an average of 79% (SD = 14%) for Brain19 and Brain20 and 65% (SD = 3%). All suggested SBS displayed great C>T dominance and SBS19 and SBS30 nearly exclusively account for this class of base substitutions. SBS19 is of unknown aetiology, SBS30 is related to base excision deficiency caused by inactivating mutations in *NTHL1*, SBS32 is caused by the immunosuppressive agent azathioprine and SBS44 is a signature for defective mismatch repair. The mutational spectra of germline mutations derived from donor-matched bulk sequencing data displayed none of the signatures suggested by SNV calling results from single cell sequencing data. Since treatment of all five neurotypical donors with azathioprine and causative mutations within all five donors necessary for SBS30 and SBS44 that were not present in the donor-matched brain bulk sequencing data were extremely unlikely, the deconvolution of mutational spectra was considered inefficient to distinguish true somatic variants from WGA-related artefacts in the present data set with substantial levels of technical noise.

Since the mutation calls from all algorithms were implied to primarily include WGA-related artefacts, a consensus set of WGS SNV calls per neuron was defined for further analyses by all base substitutions that passed the appropriate filters of at least two different algorithms. This consensus set was explored for the presence of shared mutations across multiple cells from the same donor. When restricting to variants present in at least two cells, a greatly increased burden of 5,513 and 3,048 shared mutations for Brain19 and Brain20 was observed compared to 264, 281 and 291 shared mutations in the remaining three donors. Base substitutions shared across at least two donors were strongly enriched in C>T mutations with an average of 97% for

Brain19 and Brain20 and 58% for Brain21 to Brain23. While this represented a substantial relative depletion of C>T mutations for Brain21 to Brain23, the relative fraction of C>T mutations for Brain19 and Brain20 was largely unchanged and the potential WGA batch effect in the mutational burden was still prominent. Phylogenies were reconstructed using the single cell-specific tree inference algorithm SCITE ²⁶¹. Given the high ADO rate and pattern of mutation calls that were in conflict with proper phylogenies, the consensus trees derived from SCITE output included no resolved lineage relationships and corresponded to one polytomy of all ten cells.

Restricting to SNV calls in only two neurons can still retain WGA-related artefacts that appear in the same locus as a true biological variant in another neuron from the same donor. Since this appeared to retain substantial amounts of technical artefacts in Brain19 and Brain20, a more stringent set of shared mutations was defined by requiring detection in at least three neurons from the same donor. While no germline variants that were called in donor-matched bulk sequencing data were included in the consensus WGS SNV set, variants present in at least three of ten post-mitotic neurons from the same donor were restricted to loci with coverage of at least 10 reads and VAF less than 10% in the corresponding bulk sequencing data. Given these restrictions only 8, 4, 5, 2 and 4 shared mutations could be identified for Brain19 to Brain23, respectively. Similar to the lineage reconstruction for mutations shared between at least two cells, SCITE could not resolve any phylogenetic relationships between the cells and the consensus tree corresponded to one polytomy of all ten cells per donor.

In summary, high levels of WGA-related artefacts were indicated during the SNV analyses of the genomic layer of single cortical neurons. While this was observed for all cells from all five donors, a WGA batch effect further amplified the level of technical noise for Brain19 and Brain20. Notably, the level of WGA-related artefacts did not seem to be reduced in the results from single cell-specific compared to conventional SNV calling algorithms. Since deconvolution of mutational spectra and restriction to base substitutions shared between multiple cells from the same donor also suffered from substantial levels of technical noise or could only identify very few mutations, SNV calling solely based on the genomic layers was considered to be ineffective in the present data set.

Integrated DNA-and-RNA SNV Calls Uncover a Unique Mutational Spectrum Compared to Individual Molecular Layers

For the integrated DNA-and-RNA calling approach, SNV calls from RNA-seq and the consensus WGS SNV set were overlapped and identical base substitution calls from both molecular layers were accepted into the provisional set of integrated variants. Using this approach, a total of 56 variant calls could be detected across all 50 cells. Notably, no integrated variant was detected in 24 cells and the maximum of seven integrated SNV calls was observed in a single cell from Brain19. Due to the low number of variants, it was feasible to visually confirm the robustness of variant calls. Alignment tracks from the genome and transcriptome data were visualised in JBrowse and SNV calls that were supported by noisy reads, terminal variants and in direct proximity to oligonucleotide repeats were excluded to derive the final integrated set of SNV calls (see Figure 32 for examples). After manual confirmation, 47 of the 56 initial variants were retained in a total of 24 cells (Figure 33).

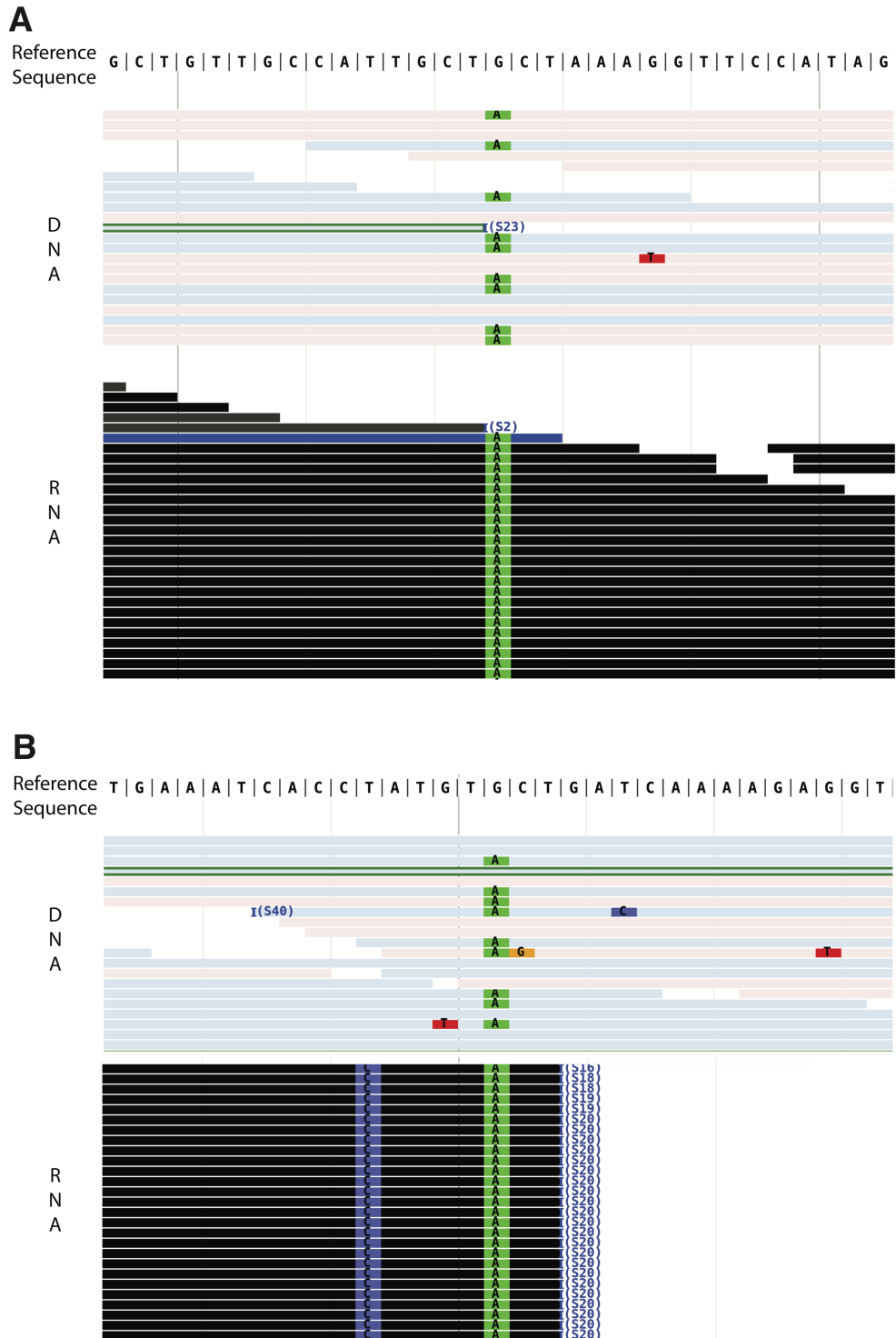


Figure 32: Quality of Integrated DNA-and-RNA SNV Calls can be Assessed by Visual Inspection of Corresponding Alignments. WGS and RNA-seq alignments were inspected on all loci with a putative DNA-and-RNA SNV call. **(A)** High quality calls are supported by variants across varying read positions and few to none additional variants on the supporting reads. **(B)** Low quality examples are found in regions with high variant density and are often supported by a terminal variant with fixed read position in RNA-seq, which implies mapping artefacts.

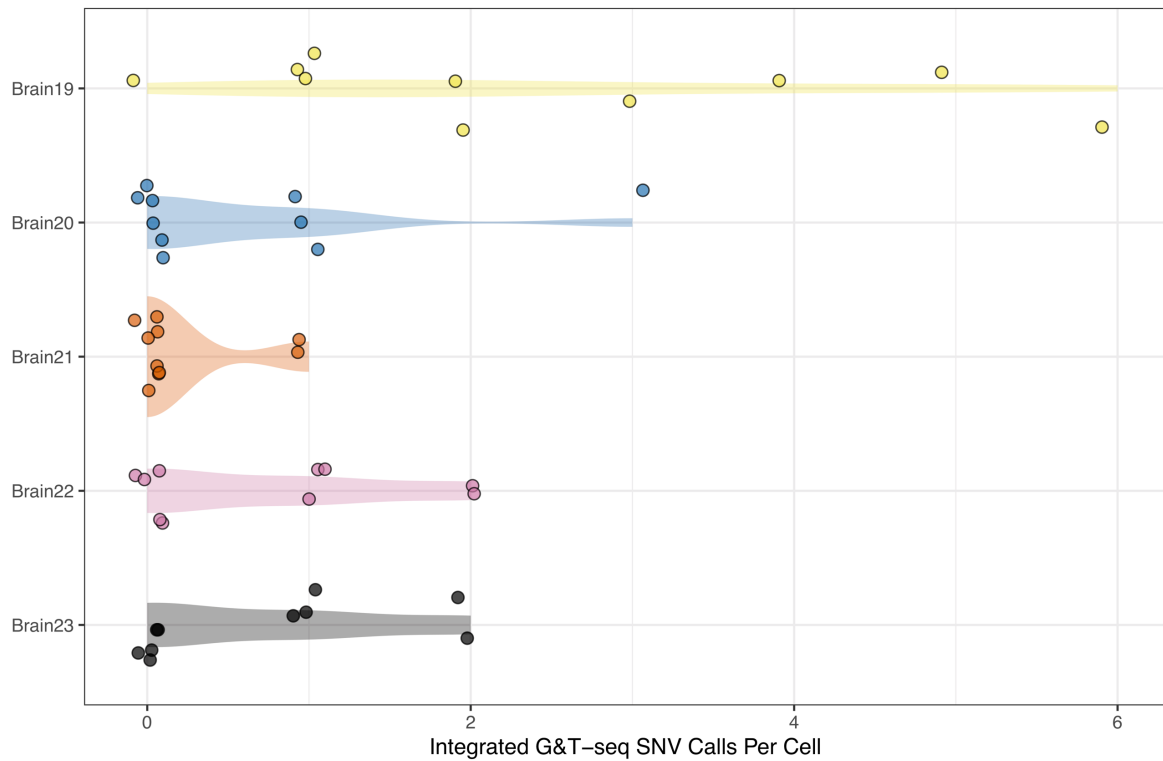


Figure 33: A Total of 47 Integrated DNA-and-RNA SNV Calls was Observed in the G&T-seq Data Set. Individual SNV calls from WGS and RNA-seq were combined and after visual inspection, 47 variants in 24 of the 50 considered neurons could be confirmed.

Since integrated variants have to be called both in RNA-seq as well as at least in two variant callers in WGS data, the fraction of the genome in which integrated SNV calls are possible, is limited to loci with genome coverage in alignments from both molecular layers. Moreover, Caveman and SCCaller apply thresholds to the minimum coverage during variant filtering, so that a minimum of eight DNA reads was necessary to include a locus into the fraction of the genome with sufficient power for DNA-and-RNA SNV calls. For the present G&T-seq data set, only 0.3% (SD = 0.1%) of the genome was sufficiently powered to call SNVs using the multiomics approach. However, the powered genome fraction significantly correlated with the number of SNVs called per cell, providing an explanation for the variable amount of zero to six integrated SNV calls per neuron (Kendall's tau = 0.4, $p < 1\text{-e}4$). Further factors with impact on the ability to generate DNA-and-RNA calls are the ADO rate in WGS, which is equivalent to a false negative rate to detect true genomic variants (Figure 29B), and the cell-specific confirmation rate of true genomic variants in the RNA-seq data (Figure 24B). Considering these factors, a somatic mutation rate was computed and extrapolated to

a genome-wide burden. While this was a very approximate extrapolation based on a limited fraction of the genome, the obtained estimates ranged between 221 and 2560 genome-wide base substitutions for all cells with integrated SNV calls (Figure 34). Notably, the substantial difference in mutational burden for Brain19 and Brain20 compared to the remaining three donors in SNV calls from WGS data was no longer present and the highest mutational burden was predicted in a cell from Brain23 (F[4,20], $p = 0.7$, one-way Anova).

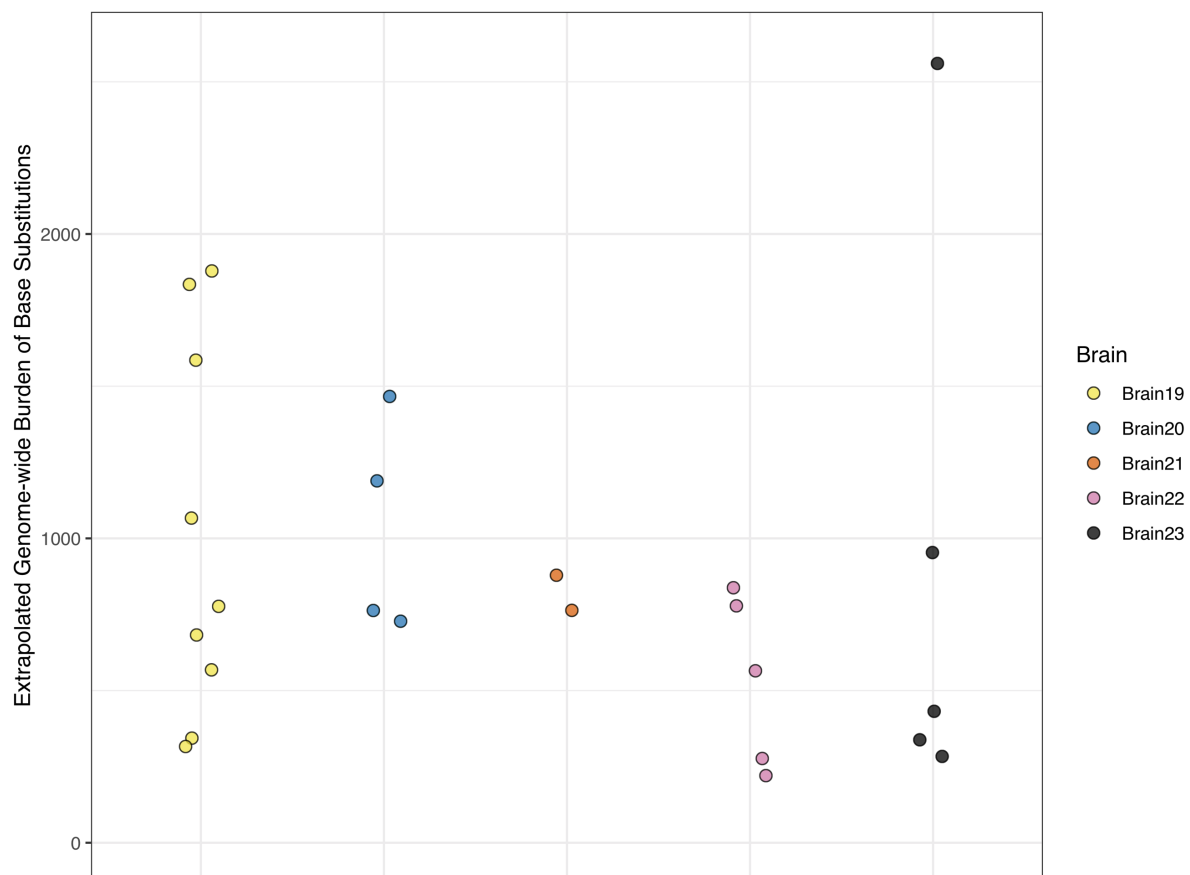


Figure 34: A Genome-wide Burden in the Hundreds to a Few Thousand Base Substitutions per Post-mitotic Neuron is Estimated Using the Integrated DNA-and-RNA SNV Calls. The genome-wide burden of base substitutions was extrapolated in cells with at least one integrated DNA-and-RNA variant based on the amount of integrated variant calls and the fraction of the genome with sufficient power to apply the integrated approach in. These results were corrected using the WGS ADO rate and the confirmation rate of genomic variants within RNA-seq data.

While a more reasonable mutational burden was predicted by the integrated SNV calling approach, the low amount of variant calls made it necessary to restrict the mutational spectrum to the six main substitution classes. The mutational spectrum from the integrated SNVs featured a unique distribution across the six main substitution classes while reducing the main source of non-genomic SNV calls in the WGS and RNA-seq data (Figure 35). C>T substitutions as main WGA-related artefact was reduced to 49% in the DNA-and-RNA compared to over 80% for WGS-based SNV calls. Compared to SNV calls from RNA-seq, the substantial amount of 46% of T>C mutations that was implied to include to A-to-I editing was reduced to 27% in the integrated SNV calls. Notably, the relative fraction of different substitution classes for the DNA-and-RNA calls was not only the average between variant calls from the individual molecular layers but was depleted compared to both layers for C>G and T>A mutations and substantially elevated for T>G substitutions.

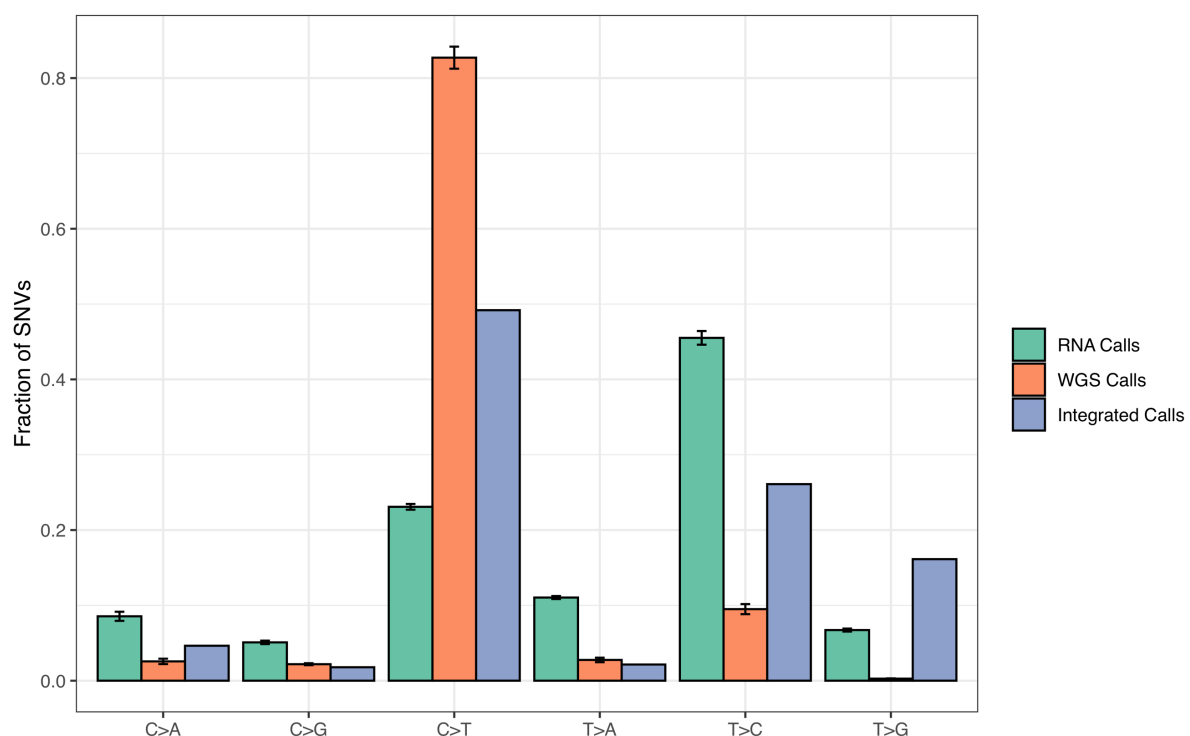


Figure 35: Integrated DNA-and-RNA SNV Calls Display a Unique Mutational Spectrum Compared to Base Substitutions in the Individual Molecular Layers. Base substitutions were classified according to the pyrimidine base change of the Watson-Crick base pair. SNV calls from RNA-seq and WGS of the same single cells display complementary spectra, caused by different sources of biological variation and technical noise for both data types. The spectrum for integrated SNV calls mitigates the dominance C>T and T>C calls from WGS and RNA-seq, respectively. Moreover, unique enrichment or depletion can be observed for T>G, C>G and T>A mutations.

Integrated DNA-and-RNA SNV Calls are Enriched in Highly Expressed Genes

The powered fraction of the genome for the multiomics SNV calls as defined in the previous result section was annotated with respect to their genomic features as defined in the Ensembl version 73 annotation. Within the powered genome background, 13% (SD = 2%) accounted for promoter, 24% (SD = 3%) for exons, 45% (SD = 3%) for introns and 18% (SD = 3%) for intergenic regions. It is important to stress that the genomic overlap and the corresponding feature annotation only indicates the breadth of genome coverage for the transcriptome layer since no hard depth filters were applied to SNV calls in RNA-seq data. Considering the total amount of RNA-seq reads, only 7% (SD = 2%) map to intergenic regions and no substantial DNA contamination was observed in RNA-seq data.

Subsequently, the loci of the 47 DNA-and-RNA SNV calls were correspondingly annotated and revealed a significant enrichment in genic regions ($p < 2e-3$, binomial test), while the distribution between promoter, exon and intron did not significantly deviate from the background distribution (Bonferroni-adjusted $p > 0.4$; see Figure 36). The genic enrichment was expected due to the integration with SNV calls in RNA-seq. Only one of 47 variants was annotated as intergenic and the substantial read support in RNA-seq for this variant either represented expression from an unannotated element or could derive from gDNA contamination during cDNA synthesis. Notably, three other neurons displayed substantial coverage in RNA-seq data at the same locus, suggesting genuine expression rather than gDNA contamination.

Cell-specific percentile ranks of expression were computed for all cells with at least one DNA-and-RNA SNV call to explore the relationship between transcriptional activity and the occurrence of base substitutions. Genes without noticeable expression were excluded to avoid inflating percentile ranks by unexpressed genes and to prevent erroneous association with zero dropouts that are common in single cell and single nucleus RNA-seq data¹⁹⁰. Given this limitation, a total of 36 genes could be associated with the 46 DNA-and-RNA SNV calls within genic regions. Genes associated with integrated SNV calls were highly expressed with an average cell-specific percentile rank of 63% (SD = 30%) and specifically enriched in the highest 3-quantile with 21 of the 36 genes ranking amongst the 33% of the most highly expressed genes per cell ($p = 2e-3$, binomial test; see Figure 37).

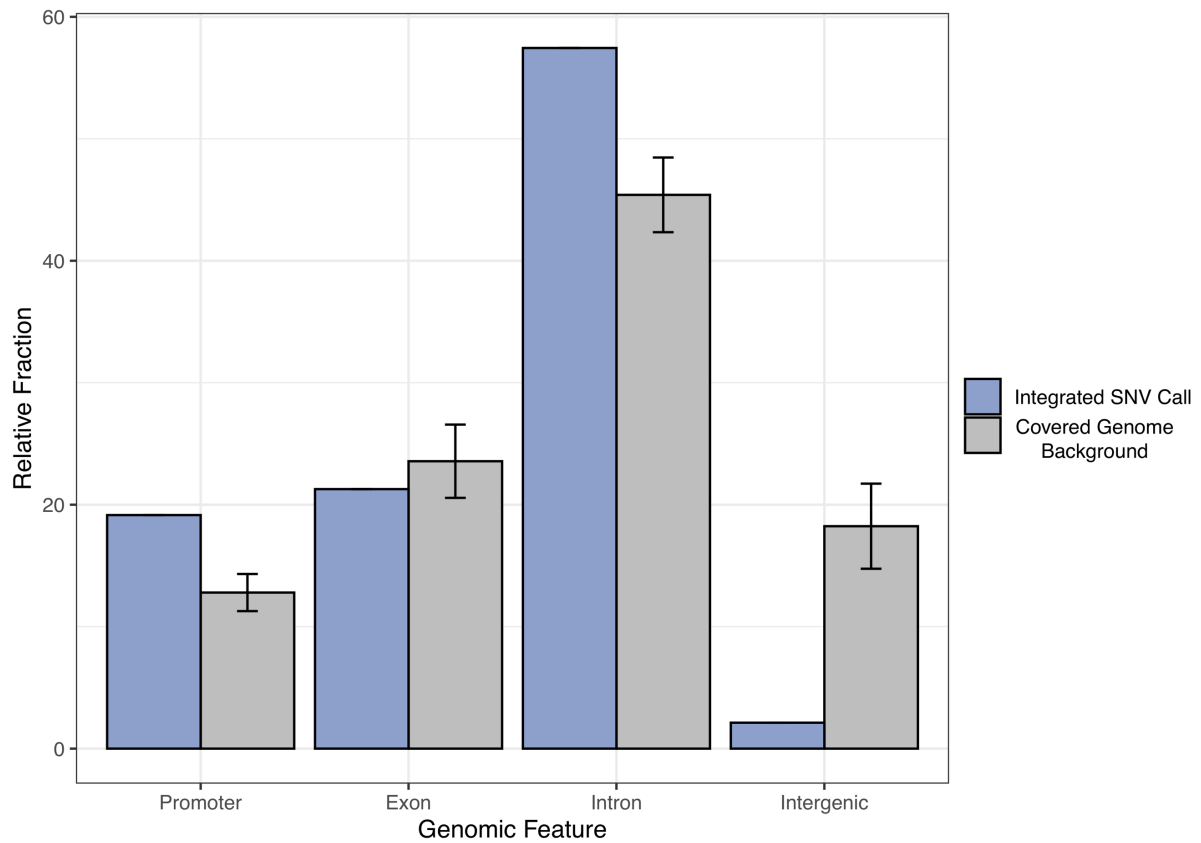


Figure 36: Integrated DNA-and-RNA SNV Calls are Evenly Distributed Across Genic Features Compared to the Genome Background with Sufficient Coverage for the Multiomics Approach. The genome background with sufficient coverage in WGS and RNA-seq as well as the loci of 46 integrated variant calls were annotated with respect to their genomic features according to the Ensembl version 73. Only one integrated variant was found within intergenic regions, which was expected due to the necessary overlap with RNA-seq. Within genic regions, the integrated calls were distributed across promoter, exon and intron regions as expected given the covered genome background.

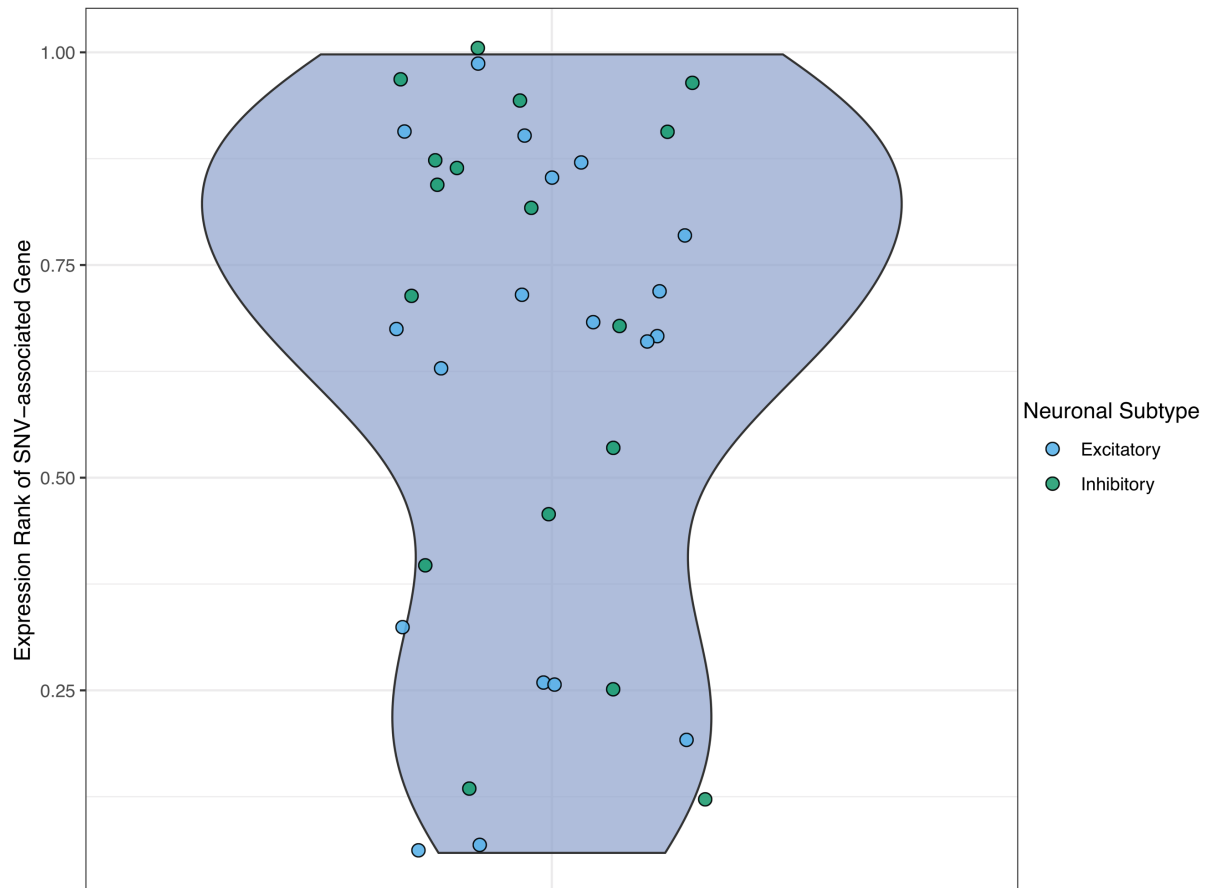


Figure 37: Integrated DNA-and-RNA Variant Calls are Enriched in Highly Expressed Genes. The cell-specific percentile ranks of gene expression were computed with restriction to genes with FPKM values of at least one. Given this restriction, 36 genes could be associated with the 46 DNA-and-RNA variant calls within genic regions. Genes that could be associated with an integrated variant call, were significantly higher expressed than expected by chance ($p = 2e-3$, binomial test for enrichment in the highest 3-quantile).

Discussion

The analysis of G&T-seq data focussed on 50 cortical neurons from five neurotypical donors was presented in this chapter. First, the transcriptome and genome of single neurons was analysed individually and similar to conventional approaches. Subsequently, information from the multiomics data was combined to derive an integrated set of DNA-and-RNA SNV calls and put into context with cell-specific transcription levels.

From the RNA-seq analysis, glutamatergic excitatory and GABAergic inhibitory neurons could be identified as the main neuronal subtypes and no robust sub-classification based on known marker genes was possible in the present data set (Figure 20 and 23). While one small cluster in addition to the two main clusters was implied by the unbiased cell type identification both in the initial shallow and the combined shallow and deep RNA-seq data set, the small clusters displayed expression of the most significant marker genes from a corresponding main cluster (Figure 20B and 23B). Notably, the small cluster identified in the initial shallow RNA-seq data set was grouped with excitatory neurons and the small cluster in the combined shallow and deep RNA-seq data set was closely related to the main inhibitory cluster (Figure 20 and 23). The most significant marker genes identified for the small clusters did not overlap with previously published marker genes to facilitate a further subtype identification within excitatory or inhibitory neurons. Previous studies could identify several of these subtypes but considered at least twice as many cells at higher read depth compared to the presented data set with 157 neurons^{269,270}. Even more comprehensive efforts profiled the transcriptome of nearly 20,000 human neuronal nuclei and thus, had much greater power to cluster single cell-derived transcriptomes at higher granularity²⁷¹. Since neuronal subtype identification was not the primary aim of this project, the cells within the small consensus clusters were grouped with their corresponding main cluster for further analyses. Notably, this affected none of the 25 excitatory neurons that were selected for WGS and only three of the 25 corresponding inhibitory neurons.

SNV calls from RNA-seq data were generated with a focus on sensitivity to maximise the probability to confirm true genomic variants with complementary RNA-seq calls

during the eventual integration. Germline SNPs from donor-matched bulk sequencing data were used as ground truth to derive the confirmation rate of genomic variants with RNA-seq coverage to be 58% on average across all cells (Figure 24A). This cell-specific confirmation rate was limited by the strongly bimodal VAF distribution across all loci in the RNA-seq data (Figure 24B). Nearly all loci displayed VAFs below 25% or above 75%, indicating high levels of allele-specific expression, stochastic bursting of gene expression or technical dropouts. These factors are known to generate strongly skewed monoallelic expression patterns in single cell RNA-seq data with mono-allelic expression being detected in up to over 76% of heterozygous germline SNPs ²⁷⁶⁻²⁷⁹. While global transcriptional profiles were shown to be similar between single cell and single nucleus RNA-seq data, the effect of transcriptional bursts might be even more pronounced in data from single nuclei ^{280,281}. This stochastic pattern of allele-detection caused an average of 91% of SNV calls from RNA-seq data to be homozygous. Since this mainly reflects biological and technical limitations of the data, RNA-seq variant calls were restricted to locus and alternative allele while information about homo- or heterozygosity was not considered in further analyses. Furthermore, the germline confirmation rate of 58% suggested a highly sensitive variant calling approach in the RNA-seq data given the stochastic detection of mono-allelic expression (Figure 24).

Germline-filtered RNA-seq variant calls were substantially enriched in T>C substitutions compared to variant calls on ERCC control sequences (Figure 26). While base substitution calls on ERCCs only represent technical artefacts, somatic SNV calls from RNA-seq data can also derive from genomic variation as well as post-transcriptional modifications such as RNA-editing. As explained in the corresponding result section, the detection of T>C substitutions is indicative of to A-to-I editing, which is the main form of canonical editing in humans ²⁷³. The amount of base substitutions corresponding to canonical A-to-I editing was significantly higher in excitatory compared to inhibitory neurons and affected a higher proportion of the expressed genes in the corresponding subtypes (Figure 27A). Genes associated with recurrent editing events were enriched in GO terms regarding calcium channel activity in both subtypes and the glutamate receptor activity in excitatory neurons (Figure 27B).

RNA editing is widespread in humans with potentially over a hundred million sites in most human genes being affected in a cell type and state-specific manner ^{273,282,283}.

Notably, RNA editing is especially prevalent in the human brain and was shown to play a crucial role in regulating mRNA trafficking and direct post-translational modification of ion channels and neurotransmitter receptors ^{273,282,283}. One of the earliest and best characterised examples of RNA editing in human brain is the modification of the calcium ion channels of the AMPA glutamate receptor ²⁸⁴. Only one published study specifically addresses RNA editing in single cell RNA-seq data thus far ²⁸⁵. The authors investigate the levels of A-to-I editing in 268 human brain cells. These cells are divided into seven subgroups including various glial cells and identify neurons as the cell type with highest prevalence of A-to-I editing. Similar to the results in this study, the authors also cannot detect a correlation between A-to-I editing with ADAR expression (Figure 28). Moreover, they observe a clear bimodal pattern of editing across all read depths, similar to the observations in the presented G&T-seq data (Figure 24B and 26B). They also observe that the vast majority of editing events is only present in a single cell and conclude that the true extent of RNA editing events is nearly impossible to detect in comparable bulk sequencing ^{283,285}. Collectively, all these findings support that substantial levels of RNA editing were detected in the SNV calls from G&T-seq transcriptome data. While considerable differences in various tissues and cell types are known for RNA editing, this is the first time that differences in the transcriptome-wide levels of A-to-I editing were observed for the main human neuronal subtypes.

The analysis of SNVs from single cell WGS data with four different variant callers implied extreme mutational burden for all selected MDA products (Figure 30). The two SNV calling algorithms Caveman and GATK established for bulk sequencing suggested a mutational burden of around 70,000 base substitutions for Brain21, Brain22 and Brain23 and around 300,000 – 400,000 mutations for Brain19 and Brain20. For comparison, all studies focussed on SNVs in normal tissues identified a burden between several hundred up to several thousand base substitutions. These studies have been performed in various tissues and used different methodological approaches including bulk sequencing of clonal cell cultures or microdissections as well as MDA-amplified single cell sequencing ^{77,84,86,87,91,92,102-104,274}. Furthermore, the corresponding mutational spectra displayed an overwhelming dominance of C>T mutations (Figure 31). Notably, C>T substitutions have been identified previously as

primary technical artefact introduced through cell lysis and amplification in conventional MDA protocols ^{104,229,275}.

These findings suggested WGA-related artefacts as the primary source for SNV calls from the single neuronal sequencing data with an additional WGA batch effect. A general trend of even further increased burden in neurons from Brain19 and Brain20 and a comparable burden between the remaining three donors was observed for all SNV callers (Figure 30). Cells from Brain19 and Brain20 already displayed lower GbS quality scores, which were shown to predict WGS quality in the previous MDA implementation chapter (Figure 18 and 21). Since the cells from Brain19 and Brain20 were processed in the same WGA batch, this increased burden is most likely to attribute to a further elevated level of technical noise introduced in this batch given the general dominance of WGA-related artefacts in the present WGS data.

Notably, it had been established that G&T-seq does not introduce elevated levels of artefacts compared to conventional processing in the previous MDA implementation chapter and the data quality for Brain21, Brain22 and Brain23 is comparable to several published data sets. Using conventional MDA approaches on single cells and nuclei was reported to result in ADO rates of over 43% compared to an average of 53% in the presented neuronal data ^{229,286}. Moreover, SNV calls from conventional algorithms were shown to result in around 32 errors per megabase ²²⁹. Given that SNVs could be called in around two thirds of the genome after excluding insufficiently covered and highly repetitive regions, this corresponds roughly to the 70,000 mutations called for Brain21 to Brain23 from Caveman and GATK.

In addition to the two conventional SNV calling algorithms, two single cell-specific callers were used that were designed to mitigate WGA-related artefacts. SCCaller estimates local non-uniformity of amplification from gHets and considers this for final variant calls and LiRA uses read-backed-phasing of gHets with potential somatic variants ^{104,206}. The genome-wide burden from SCCaller were the most conservative ones from all four algorithms with about 20,000 mutations for Brain21 to Brain23 and over 100,000 for Brain19 and Brain20 (Figure 30). In the original SCCaller publication, nearly 23,000 mutations with around 90% C>T substitutions are detected in WGS data from conventionally MDA-amplified cells. Remarkably, less than 1,000 SNVs are

called with a drastically reduced fraction of C>T substitutions in cells that undergo a modified MDA approach where cell lysis and DNA denaturation are performed on ice¹⁰⁴. These results are in line with the observations for cortical neurons from Brain21 to Brain23 in this data set and confirm that technical artefacts are the primary source of SNV calls in conventionally MDA-processed cells even when single cell-specific variant callers are used (Figure 30 and 31).

The genome-wide burden estimated by LiRA was around 45,000 for Brain21 to Brain23 and over 400,000 - 600,000 mutations for Brain20 and Brain19, respectively. While the suggested genome-wide burden for Brain21 to Brain23 was lower compared to the estimates of around 70,000 from conventional SNV callers, the estimates in Brain19 and Brain20 with higher levels of technical noise even surpassed the 300,00 - 400,000 mutations called with the conventional algorithms. This suggests that WGA-related errors are not completely accounted for in LiRA, especially in high artefact backgrounds. One of the main problems inside LiRA is the assumed binomial sampling from both strands of the original DNA molecule. This even sampling is considered when deciding for sufficiently powered loci to call phasing-backed SNVs²⁰⁶. However, any DNA molecule with artefacts introduced during cell lysis or early in the MDA amplification process that is affected by a sufficient level of uneven amplification will give rise to sequencing data that can be in perfect phasing agreement. The ADO rates of over 50% observed in the presented data set provide a direct estimate of the extent of non-uniform amplification between homologous loci (Figure 29B). While the degree of non-uniformity between the Watson and Crick strand from the same original molecule has not directly been assessed so far, it is not unreasonable to assume a similar level to the one observed between homologous loci. Consequently, current computational methods are likely to retain a fraction of WGA-related artefacts, even in modified MDA approaches with reduced technical noise.

Subsequent to the analysis of the individual molecular layers from G&T-seq, SNV calls from RNA-seq and WGS of the same neurons was integrated to derive a set of DNA-and-RNA SNV calls. After manual inspection, a total of 47 integrated variants was confirmed in 24 of the 50 neurons (Figure 32 and 33). Due to the requirement for sufficient coverage in RNA-seq and WGS data and exclusion of repetitive regions, only a small fraction of less than 1% of the genome was sufficiently powered to derive

integrated calls. Notably, the amount of DNA-and-RNA SNV calls per cell significantly correlated with the corresponding fraction of sufficiently powered genome. Therefore, a very rough extrapolation under consideration of false positive rates in WGS and RNA-seq data was performed (Figure 34). This extrapolation has several limitations, most notably that no mutations could be confirmed in about half of the cells and the extrapolation is based on expressed somatic variants that are functionally different from the majority of the genome. Despite these limitations, the corresponding estimated burden of several hundred up to about 2,500 mutations agree with the burden observed in other normal tissues as well with previous publications investigating somatic SNVs in human neurons ^{77,84,86,87,91,92,102-104,274}.

In addition to the more reasonable estimated burden from integrated DNA-and-RNA SNV calls, they also displayed a unique mutational burden compared to the individual molecular layers (Figure 35). As mentioned earlier, the predominant sources of variation were over 80% C>T substitutions as likely WGA-related artefacts in the WGS data and about 46% T>C substitutions in RNA-seq that included a high proportion of A-to-I RNA editing ^{104,229,273,275}. The integrated SNV calls displayed drastically reduced relative contributions of these classes with 49% accounting for C>T substitutions and 27% for T>C substitutions. While these substitution classes still account for most of the integrated SNV calls, this is in line with observations from WGS of other normal human tissues that have not undergone single cell sequencing ^{84,86,91,92}.

The burden and mutational spectra for human cortical neurons have been analysed in two previous single cell studies from the same research lab using WGS from MDA-amplified products ^{102,103}. In both studies, the authors apply a modified MDA approach similar to the one introduced in the SCCaller publication ¹⁰²⁻¹⁰⁴. In their first analysis, the authors use the consensus calls from three conventional callers and error corrections based on mutation calls of the hemizygous X chromosome in men. They conclude that around 1,500 base substitutions were present in neurons from two teenagers and a 42-year-old individual. Notably, the mutational spectrum displayed a similar fraction of around 80% C>T substitutions as observed for SNV calls from conventional callers in the present data set (Figure 31) ¹⁰³. In their follow-up study, the authors acknowledged the possibility of MDA artefacts and applied LiRA to establish the burden and base substitution spectrum in human post-mitotic neurons ¹⁰². Their

main findings in neurons from healthy donors was an age-dependent mutational burden of less than 1,000 SNVs for new-borns and teenagers that increases to around 3,000 – 4,000 substitutions for individuals over 80 years of age. Furthermore, they identify two signatures that primarily contribute to the mutational burden in healthy donors. One signature is similar to the clock-like signature 5, while the second signature is overwhelmingly C>T dominant and virtually identical to the signature derived from LiRA in the presented G&T-seq data with the exception of a small contribution of T>A substitutions (Figure 31) ^{66,102}. Notably, the C>T dominant signature accounted for nearly 100% of the 300 – 900 mutations detected in infants. The authors argue that this is in great alignment with the 200 – 400 somatic base substitutions reported for neural progenitor cells up to 20 weeks of gestation ^{87,102}. However, the mutations detected in sequencing data from clonal cultures of neural progenitors were predominantly C>A mutations with less than 30% accounting for C>T substitutions ⁸⁷. Notably, cell culture has been demonstrated to introduce high levels of C>A mutations and the significant enrichment of C>A mutations in private compared to mosaic mutations, suggests a complementary source of artefacts in the results for the neural progenitor cells ^{79,87,287}.

While estimates for the mutational burden from the integrated DNA-and-RNA approach are too approximate to allow a detailed comparison, the average estimated burden of less than 1000 SNVs in cells with at least one integrated call is nominally less than the burden of usually around 1000 – 2000 base substitutions assigned by LiRA in the most recent study exploring the SNV landscape in adult human neurons (Figure 34) ¹⁰². Moreover, the average spectrum derived by LiRA in modified MDA approaches with less technical noise are still enriched in C>T substitutions with around 55% compared to the relative fraction of 49% for the integrated DNA-and-RNA approach (Figure 35) ^{102,206}. Collectively, the critical analysis of published data together with the results presented in this chapter, suggest that a remaining fraction of MDA artefacts obscure some of the previously published insights into the mutational landscape in post-mitotic human neurons. However, the present data set with 50 cortical neurons and high levels of technical noise is too limited to finally resolve all discrepancies with sufficient granularity.

To improve the proposed method, the G&T-seq workflow should be adjusted to conform with the modified MDA approaches that have been demonstrated to significantly reduce the level of technical noise instead of following the conventional MDA guidelines ¹⁰²⁻¹⁰⁴. Furthermore, application to a greater number of neurons from several donors across a wide age range would allow for greater power for mutation detection and the analysis of age-related effects. More importantly, improving the fraction of genome coverage with sufficient power for the integrated DNA-and-RNA approach would facilitate the detection of a greater number of variants and thus, allow for a more detailed mutation spectrum analysis and a more robust genome-wide extrapolation. While the modified MDA approach potentially improves genome coverage and uniformity for WGS data, the RNA-seq data represents the more limiting factor. Currently, G&T-seq is restricted to analysis of polyadenylated mRNA and the Smart-seq2 protocol used in the present data set allows for the greatest detection of full-length transcripts of currently available protocols for single cell RNA-seq ^{250,288,289}. Therefore, improving genome coverage for the transcriptome layer of G&T-seq is unlikely given the current methodological limitations.

While the presented multiomics approach still has limitations and cannot fully resolve outstanding inaccuracies from previously published studies, the parallel acquisition of genome and transcriptome of the same cell could directly address the suggested role of active transcription for mutagenesis in post-mitotic neurons. The confident set of DNA-and-RNA variant calls were significantly enriched in highly expressed genes (Figure 37). Notably, this association was directly derived from cell-specific expression values and confirmed a previous suggestion based on correlation to expression values from databases ¹⁰³. Remarkably, high transcription is correlated with low mutational frequencies in cancer genomes ²⁶⁷. The positive correlation with active transcription previously indicated and now confirmed in the presented multiomics data, points out a unique mutational characteristic in post-mitotic neurons compared to highly dividing cells.

In summary, the presented G&T-seq approach provides a method with unique benefits to SNV calling in single cells. Despite limitations in genome coverage and a low number of high-confidence variants, a refined mutational spectrum for base substitutions and a direct role of high transcriptional activity for mutagenesis could be established for human cortical neurons.

Mutational Processes and Clonal Dynamics in Prostatic Epithelium

Introduction

This chapter discusses mutational processes and burden in histologically normal prostatic glandular epithelium and characterises clonal dynamics within ductal networks of the human prostate. Targeted microdissections were subjected to WGS to infer the mutational status and signatures in histologically normal prostatic epithelium. Additionally, long-ranging ductal networks were reconstructed for one donor to overlay WGS samples with spatial information and their corresponding morphological relationship.

A brief summary of prostate anatomy and development, relevant aspects of benign prostate hyperplasia (BPH) and prostate cancer as well as previous work addressing stem cells and clonal dynamics within the normal human prostate are presented as background for this chapter.

Anatomy and Development of the Human Prostate

The human prostate is an accessory gland of the male reproductive system, which is situated directly underneath the bladder and surrounds the prostatic urethra as well as the ejaculatory ducts²⁹⁰. The human prostate is classically divided into four different zones. Glandular epithelial tissue is distributed throughout the central, the transitional as well as the peripheral zone while the anterior fibromuscular stroma (AFMS), representing the fourth zone, contains no ductal tissue²⁹¹. The central zone surrounds the ejaculatory ducts while the transitional zone surrounds the proximal urethra approximately until the verumontanum, where the ejaculatory ducts enter the prostatic urethra. The peripheral zone covers the posterior and lateral prostate and largely surrounds the central and transitional zones²⁹¹. Within the normal prostate in younger men, the central zone forms around 25% of the glandular tissue, the transitional zone accounts for 5% and the peripheral zone for 70% (Figure 38)²⁹².

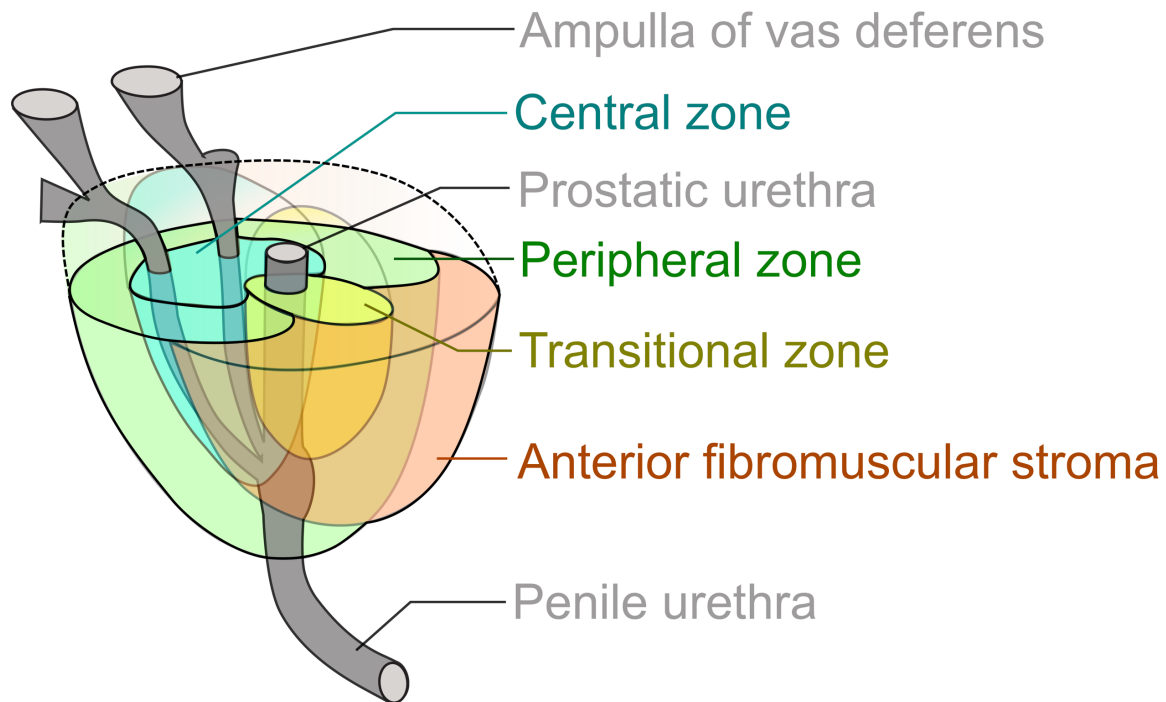


Figure 38: Anatomical Zones of Human Prostate. The human prostate can be divided into four anatomical zones. Only the central, transitional and peripheral zone contain glandular epithelium (This figure is a derivative of “Creative Commons Zones of the prostate” by Mikael Häggström licensed under CC0 1.0).

The glandular function of the prostate is performed by a network of epithelial compound tubular-alveolar glands that are embedded in the prostatic stroma²⁹³. Approximately 25 – 30 independent glandular subunits extend from the urethra in a complex branching pattern before terminating in multiple acini^{294,295}. The glandular tissue mainly consists of a two-layered stratified columnar epithelium of luminal and basal cells. In the human prostate, luminal and basal cells contribute evenly to the glandular epithelium²⁹⁶. Traditionally, only luminal and basal as well as very rare neuroendocrine (NE) cells were described in prostatic epithelium based on cellular locations as well as a selected subset of surface antigens and marker gene expression²⁹⁷. More recently, an unbiased approach applying single cell RNA-seq to histologically normal prostatic epithelium, could confirm luminal, basal and NE cells but also identified two further epithelial cell types that were enriched in the urethra and peri-urethral ducts²⁹⁸.

The complex branching pattern of the epithelial glandular subunits is established during embryonic and extended during pubertal development^{290,299,300}. Development of the human prostate can be divided into four stages²⁹⁰. The first stage is the

androgen-induced fate determination that results in prostate development in males from the pelvic part of the urogenital sinus (UGS) at around 8 – 9 weeks of gestation. Shortly thereafter, prostatic buds arise from the UGS epithelium that elongate into the surrounding UGS mesenchyme. During this stage, a primitive ductal network composed of solid epithelial cords is established. Starting at around 11 weeks of gestation, more pronounced bud elongation and complex branching morphogenesis can be observed within the glandular network ^{290,301}. While some anatomical trajectories are conserved and paracrine signalling between the UGS epithelium and mesenchyme is essential for budding and branching morphogenesis, it is thought that branching follows a stochastic process, where budding tips explore the surrounding mesenchyme according to a model described by branching and annihilating random walks ^{290,301,302}. Repeated cycles of branching and ductal elongation driven by budding tips terminate independently when a budding tip comes into proximity of a previously established duct. This explains the complex organisation and irregular length of ducts within the same glandular subunit ³⁰². Budding tips are most prevalent during embryonic development, are still present during adolescence but are not present in adult human prostates ³⁰⁰. The final stage of development within prostatic glandular subunits comprises the canalization of the solid epithelial cords and differentiation into the mature stratified columnar epithelium. Similar to the branching morphogenesis, this process follows a proximal to distal direction within each glandular subunit ^{290,301}.

While a period of functional maturity of the prostatic glandular epithelium is observed at latest at 20 weeks of gestation, the prostate gland enters a quiescent phase in the last trimester. With the strongly increased androgen production during puberty, another substantial period of human prostate development is observed resulting in the final complex glandular architecture and full maturity ^{299,300,303,304}. The extent of glandular morphogenesis and additional branching is substantial during puberty with the epithelial to stromal ratio increasing by a factor of four in a study analysing fifty normal prostates between 0 – 30 years of age with a distinct step up in the ratio observed at around 15 years of age ³⁰⁵.

Following embryonic and pubertal development, neoformation of glandular tissue is only observed in context of the two prostate diseases BPH and prostate cancer ³⁰¹. Relevant aspects of these diseases are briefly summarised in the next section.

Benign Prostate Hyperplasia and Prostate Cancer

BPH and prostate cancer are both common diseases in older men ^{306,307}. Since the age profile, several symptoms and risk factors as well as features like neoformation of glandular tissue are shared between these diseases, there has been controversy about whether the diseases share pathogenic mechanisms ³⁰⁸⁻³¹⁰. While the most recent meta-analysis of previous studies addressing this question concludes that BPH is associated with increased risk of prostate cancer, it is not considered a direct precursor lesion ^{293,309}.

BPH is the benign growth of nodules of glandular and stromal tissue nearly exclusively in the transitional zone. Commonly, the transitional zone accounts for over 50% of prostatic volume in men over 70 years of age compared to 5% during early adulthood ³⁰⁶. While the aetiology of BPH is not well understood, comorbidities such as type 2 diabetes and obesity are risk factors, perhaps acting through hormone activity alteration and pro-inflammatory pathways ^{293,306}. Moreover, the declining ratio of testosterone to oestradiol as well as the accumulation of dihydrotestosterone (DHT) may play a role in the disease progression of BPH. Notably, symptomatic BPH can be treated by inhibition of the enzyme 5 α -reductase, which converts testosterone into DHT ^{293,306}.

Prostate cancer is the most common solid malignancy in men ³⁰⁷. With respect to the different anatomical zones, only very few prostate cancers originate in the central zone, while the transitional and peripheral zone account for around 25% and 70%, respectively ²⁹². Around 90% – 95% of prostate cancers are acinar adenocarcinoma, which originate in the acini of glandular subunits. The most common histological variant are ductal adenocarcinoma that account for around 3% of prostate cancer cases and often present as mixed ductal-acinar carcinomas ³¹¹. Therefore, nearly all prostate cancers derive from glandular tissue, preferentially in the periphery of the complex ductal networks. Remarkably, prostate cancer is frequently multifocal and several studies support an independent clonal origin for these foci ³¹²⁻³¹⁵. However, some somatic mutations found within the cancer foci were also observed in spatially

distinct, histologically normal prostate tissue and one report proposes a common clonal origin for multifocal prostate cancer^{315,316}.

The prostate cancer genome is characterised by a modest mutational burden of around one mutation per megabase and usually only contains few focal CNVs³⁰⁷. The most frequently mutated genes comprise *TP53*, *SPOP*, *KMT2C*, *KMT2D* and *FOXA1*. A recent study of over 1,000 prostate cancers identified a total of 97 genes to be significantly mutated. Notably, great inter-individual heterogeneity exists and most of these genes are only mutated in fewer than 3% of patients³¹⁷. In contrast to the low mutational burden, prostate cancer often presents with a high amount of chromosomal rearrangements³⁰⁷. These rearrangements frequently include gene fusion with the ETS transcription factor family, which are present in over 50% of all cases of prostatic adenocarcinoma^{318,319}. In comparison to the primary disease, metastatic prostate cancer presents with significantly increased mutational burden as well as substantially reduced genomic heterogeneity, highlighting the importance of continuing evolutionary processes^{307,320,321}.

The high frequency of prostatic diseases with neoformation of glandular tissue, a highly diverse but comparably quiescent mutational landscape as well as the multifocality of prostate cancer prompt questions about somatic evolution and clonal dynamics within the normal prostatic epithelium. An overview about previous insights into these clonal dynamics and corresponding stem cell biology in the human prostate is given below.

Stem Cell Biology in the Normal Human Prostate

Most insights into stem cell biology of the prostate have been gained from rodent models. However, different anatomic organisation, shortened postnatal development as well as the absence of naturally occurring prostate cancer in these rodent models point out that stem cell organisation and clonal dynamics are likely to be substantially different between the rodent and the human prostate^{293,295}.

During embryonic development, it is well established that ducts and acini form in proximal to distal direction through elongation and branching of budding tips. Epithelial cells within these budding tips therefore contain the embryonic stem cell population

that is responsible for directed morphogenesis ^{290,301}. Similar budding tips are observed before and during adolescence but are absent in the adult prostate ³⁰⁰.

Following these two phases of development, the prostate enters the state of adult tissue maintenance. Tissue turnover and regeneration that is maintained over decades is commonly attributed to adult stem cells, which are long-lived, possess self-renewal and high proliferative capacity as well as multipotent lineage potential ³²². Due to these features, adult stem cells (ASCs) that escaped tissue-typical regulation are hypothesised to give rise to cancer and the presence of these so-called cancer stem cells has been identified in prostate cancer ³²³. Therefore, an improved understanding of normal clonal dynamics will contribute to the understanding of prostate diseases.

Some ASC compartments are large and highly proliferative. For example, the hematopoietic system has to replace up to billions of highly specialised cells and 50,000 – 200,000 stem cells are estimated to contribute to leukocyte generation alone ⁸⁶. In contrast, the normal prostate is relatively quiescent with an estimated turnover of epithelial cells in the range of 200 days ³²⁴. The existence of ASCs throughout the human prostate has primarily relied on the identification of cell populations with stem cell markers and their proliferative capacity *in-vitro* and in xenograft models ³²⁴⁻³²⁹. Collectively, these studies propose a small population of cells with stem cell-like properties scattered throughout the glandular epithelium. This stem cell-like population is primarily located in the basal cell compartment and hypothesised to account for about 1% – 2% of the cellular mass ³²⁴⁻³²⁹. Moreover, bulk RNA-seq of the basal and luminal cells confirmed stem cell-like features in the basal compartment ³³⁰.

Additionally, three studies have performed *in-situ* lineage tracking in glandular epithelium from serial sections of the normal human prostate. All of these studies relied on visual tracking of a mitochondrial DNA mutation in the cytochrome c oxidase (COX) enzyme, that can be stained using immunohistochemistry approaches ^{295,331,332}. The two earlier studies established the common clonal origin for basal, luminal and NE cells throughout the ductal system. While most terminal acini were found to be polyclonal, demonstrated by co-existence of normal and COX-deficient tissue, rare monoclonal conversion was observed indicating that single stem cells can generate

whole terminal branches ^{331,332}. The more recent study performed long-ranging reconstruction of individual glandular subunits and was able to demonstrate that some COX-deficient clones spanned the whole length from urethra-proximal ducts to distal acini. Additional functional assays established that proliferative activity was mainly restricted to the basal cell layer and decreased in proximal to distal direction. Since a higher proportion of long-ranging COX-deficient clones was detected in glandular tissue from older patients, continuous migration of COX-deficient clones was postulated and integrated into a stem cell model. According to this model, the stem cell niche is confined to the urethra-proximal ducts, where basal stem cells extend in directed streams with decreasing proliferative potential and give rise to luminal cells ²⁹⁵. In support of a specific stem cell population in the peri-urethral ducts, recent single cell RNA-seq discovered two new epithelial cell types enriched in this area that display expression similarities with immature cell types in other human tissues and prostate cancer. While their cellular function is not validated, the authors hypothesise that these cell types could be enriched for prostate ASCs ²⁹⁸.

While the *in-situ* lineage tracking studies significantly contributed to the understanding of clonal dynamics within the prostate, a unique stem cell niche in the peri-urethral ducts with distinct proximal to distal gradient of stemness contradicts the earlier findings of scattered and isolated stem cell populations throughout the basal compartment of glandular subunits ³²⁴⁻³²⁸. Moreover, COX-deficient lineage tracking relies on the existence of a single mutation that cannot be temporally dated to different phases of development or tissue homeostasis. Similarly, no subclonal relationships can be distinguished within COX-deficient clones. Here, glandular reconstruction is combined with WGS to explore the more complex mutational processes on the whole-genome-scale and correspondingly analyse the clonal dynamics within individual glandular subunits.

Methods

Prostate Samples

A snap-frozen whole prostate sample from a 59-year-old donor suitable for the 3D reconstruction and phylogenetic analysis of long-ranging glandular subunits was provided by Rakesh Heer from Newcastle University. Additional prostate samples were obtained to explore the mutational landscape and to infer the relationship between mutation accumulation and aging in normal prostate epithelium. Another snap-frozen prostate biopsy from a 70-year-old donor was provided by Rakesh Heer. Four snap-frozen post-mortem prostate biopsies from 22-, 30-, 47- and 71-year-old donors were obtained from AMS Biotechnology. Furthermore, WGS data of microdissections from normal prostate epithelium from a 59- and 78-year-old donor were made available by Luiza Moore. All samples were obtained, stored and processed with appropriate ethical approval.

Sample Preparation for Laser-Capture Microscopy

The frozen prostate samples were equilibrated to 0 °C before fixation in PAXgene Tissue Fix and the corresponding PAXgene Tissue Stabilizer (PreAnalytiX). Subsequently, the whole prostate sample was cut longitudinally into two 10 mm blocks. The prostate biopsies were of variable thickness between 2 – 5 mm and could directly be subjected to paraffin embedding. Paraffin embedding was performed by Yvette Hooks using a Sakura Tissue Tek VIP 6 Processor with standard histological tissue processing protocols. The paraffin-embedded prostate blocks were sectioned by Yvette Hooks to derive 10µm thick cross-sections of the whole prostate or biopsy region and mounted onto PEN-membrane slides (Leica). Individual sections were stained with haematoxylin and eosin (H&E) by sequential immersion into: the xylene-substitute Neo-Clear (two minutes, twice), ethanol (70%, 1 minute, twice), ethanol (100%, 1 minute, twice), deionised water (1 minute, once), Gill's haematoxylin (15 seconds), tap water (20 seconds, twice), eosin (6 seconds, once), tap water (15 seconds, once), ethanol (70%, 15 seconds, twice), ethanol (100%, 30 seconds, twice), and Neo-Clear (15 seconds, twice). The stained sections were scanned using a NanoZoomer S60 (Hamamatsu) up to a 20- or 40-fold magnification. No cancerous lesions were detected in the regions of interest that were subjected to Laser-Capture Microscopy (LCM) and WGS.

Reconstruction of Prostatic Glandular Subunits

Prostatic glandular subunits were only reconstructed for the 59-year-old donor of the whole prostate. Low-resolution images at 10x magnification were extracted in jpg format from the high-resolution digital slide scans to be able to perform image registration on 671 sequential whole prostate images. Intensity-based automatic rigid image registration was performed within Matlab. The aligned images were screened for long-ranging prostatic glandular subunits and series of several hundred image files containing the ductal trees of interest were imported into the serial section microscopy editor Reconstruct³³³. Branching of prostatic ductal trees was followed in proximal to distal fashion and regions of interest were annotated for subsequent LCM microdissection. Three-dimensional structures of LCM microdissections and their histological connection through the ductal network were visualised in R using the plotly package³³⁴. Two-dimensional projections were generated using metric multidimensional scaling (MDS) in R.

Laser-Capture Microscopy and Whole-Genome Sequencing

Prostatic ducts and acini of interest were dissected using a LMD7 laser-capture microscope (Leica) and collected into separate wells of a 96-well plate. Regions of interest were cut from up to three adjacent sections to yield an approximate total of 200 – 2000 cells epithelial cells per well. For the 59-year-old-donor of the whole prostate, most LCM microdissections were derived from the reconstructed glandular subunits. Ductal and acinar regions of interest from the remaining donors were selected only based on benign histological appearance and appropriate cell number for microdissection without detailed knowledge about their morphological relationship.

The DNA was extracted using the Arcturus PicoPure DNA Extraction Kit (Thermo Fisher Scientific) according to the manufacturer's instructions. The 96-well plates with extracted DNA were submitted to the Sanger Pipelines for bespoke sequencing library preparation from low DNA input samples and subsequent WGS. Typically, six to eight samples were multiplexed and sequenced on the same total number of Illumina HiSeq X lanes to generate 150 bp paired-end reads.

Genome Alignment and Variant Calling

The WGS data was aligned against the GRCh37 genome by the CASM Core Informatics Processing Pipeline at the Wellcome Sanger Institute. Base substitutions were called using Caveman against a donor-matched stromal prostate sample with corresponding filtering as described in the chapter presenting the G&T-seq analysis of human cortical neurons. Structural variations were called using BRASS ³³⁵.

Inference of the Relationship Between Mutation Accumulation and Aging

To infer the relationship between mutation accumulation and aging, all WGS samples derived from LCM of normal prostate epithelium that were already available within the Wellcome Sanger Institute or generated for this study were considered.

To obtain more accurate estimates for the amount of base substitutions present in every section considered, clonality and sequencing depth of the corresponding WGS data had to be considered since polyclonality or low sequencing depth will result in low estimates of the mutational burden. Sequencing depth was computed using samtools ²²³. The degree of polyclonality was estimated by the median VAF of all passed base substitution calls in a sample. Since a minimum of four variant reads was required during filtering, robust estimation of polyclonality c was only possible in samples with appropriate sequencing depth d with the following relationship:

$$d \geq \frac{4}{c}$$

For example, perfectly clonal diploid samples would display a median VAF and corresponding estimate of $c = 0.5$ and thus, only required a minimum depth $d = 8$, while highly polyclonal samples with $c = 0.1$ required a sequencing depth of at least $d = 40$. Only samples that displayed sufficient sequencing depth for their corresponding clonality estimates in at least 75% of the genome, were considered in the final regression analysis.

For these samples, the sensitivity for SNV calls based on clonality and sequencing depth was computed using a generalised linear model (GLM) provided by Simon Brunner. The GLM parameters were calculated based on LCM-derived WGS data of human liver tissue. Directly adjacent tissue sections were assumed to comprise the

same variants and thus, any diverging SNV calls in directly adjacent sections were attributed to polyclonality and insufficient sequencing depth. Subsequent to sensitivity estimation s , the mutational burden as estimated from Caveman and corresponding filtering b_r was corrected to yield the mutational burden b_c :

$$b_c = b_r / s$$

To account for dependent multiple samples from the same donor as well as the observed distribution of mutation count data, a generalised linear mixed-effect model with negative binomial response distribution and log-link function was fitted using the Laplace approximation to infer the effect of age a on the mutational burden b_c using the R package lmerTest with following formula ³³⁶:

$$b_c \sim \log(a) + (1 | Donor)$$

A linear relationship between donor age a and corrected mutational burden b_c is considered using this approach. Since a log-link function is used, the logarithm of age a is used to account for the exponentiation during computation. A random intercept per donor is considered to account for the interdependency of samples from the same donor and is indicated by $(1 | Donor)$ according to the R syntax. The age effect was derived from this model using the R package effects ³³⁷.

Mutational Spectra and Signature Extraction

Mutational spectra for base substitutions, corresponding signature extraction and comparison to SBS signatures from the ICGC PCAWG Platinum release was performed using SigProfiler v.2.1 as described in more detail in the chapter presenting the G&T-seq analysis of human cortical neurons ³⁶.

Detection of Positive Selection and Driver Mutations

Positive selection of mutations was evaluated using a dN/dS approach within the R package dndscv ³³⁸. This method compares the amount of non-synonymous to synonymous mutations within coding regions and accounts for local mutation rates as well as for the globally observed trinucleotide context of mutations. An excess of non-synonymous compared to synonymous mutations relative to what is expected by

chance indicates positive selection, while a corresponding depletion would suggest negative selection ¹¹⁷.

In addition to the unbiased detection by the dN/dS approach, genic base substitutions resulting in missense, nonsense or essential splice site mutations were intersected with the most comprehensive catalogue of base substitution driver mutations in prostate cancer comprising 97 significantly mutated genes (SMGs) as published by Armenia and colleagues in 2018 ³¹⁷. The observed change was annotated as a known driver mutation if it was present within the discovery set of Armenia and colleagues or within hotspots as annotated by the COSMIC Cancer Gene Census or the IntOGen database ^{127,317,339}.

Telomere Length Estimation

Telomere lengths were estimated from genomic alignments using the Telomerecat bam2length command with 100 iterations for estimating the impact of insert size distribution and default parameters otherwise ³⁴⁰.

Lineage Tree Inference

Base substitution profiles from LCM microdissections of reconstructed ductal subunits were used to infer their phylogenetic relationships. Adhering to the principle of the infinite site model of molecular evolution, it was assumed that only one independent mutation occurred per locus and was subsequently maintained in all cellular progeny ⁴⁰. For perfectly clonal samples, binary mutation calls can directly be used to infer the corresponding lineage tree. However, most microdissection samples were found to be oligo- or polyclonal. Therefore, not all mutations called in one microdissection derive from the same most recent common ancestor (MRCA).

The VAF distribution of base substitutions within one sample provides information about the clonal origin of these mutations. Discrete clusters of mutations observed at similar VAFs suggest a common clonal ancestor for all of these mutations. To use VAF instead of binary mutation calls for the lineage tree inference, the total read depth and the alternative allele count were obtained for all loci that were called by Caveman in at least one sample from the same prostatic ductal subunit. Formal inference of discrete clusters of mutations based on their VAFs was performed with an

n-dimensional hierarchical Dirichlet process (n-HDP). The stick-breaking process of the n-HDP results in highly concentrated clusters of mutations that display similar VAF across samples. The algorithm was implemented by Peter Campbell as previously described¹⁸. Every cluster derived by the n-HDP represents an ancestral cell that was the MRCA of all cells carrying these mutations. Under the assumption that most normal epithelial cells are diploid, the median VAF m from all mutations associated with one n-HDP-derived cluster hdp directly translates into the cellular contribution cc per microdissection i :

$$cc_{hdp,i} = m_{hdp,i} * 2$$

For example, if the median VAF for one n-HDP-derived cluster corresponds to 0.5 in one microdissection, 100% of the cells in this microdissection share the corresponding ancestral cell as their MRCA.

For the phylogenetic inference, only clusters with at least 10 mutations were considered. Furthermore, up to 20 mutations from each cluster were manually inspected to identify clusters that comprised common sequencing and alignment artefacts. Using the relationship between median VAF and cellular contribution per microdissection, lineage relationships can be inferred based on the pigeonhole principle. Since the sum of cellular contributions cannot exceed 100% per microdissection, clones providing smaller cellular contributions must be subclones to n-HDP-derived clusters with greater cellular contributions. Absolute certainty for lineage nesting was assumed if two n-HDP-derived clusters exceeded 95% cellular contribution in at least one microdissection. Furthermore, lineage nesting was strongly suggested if two n-HDP-derived clusters displayed the same pattern of relative cellular contribution in all microdissections and their sum of cellular contribution was at least 75% in at least one microdissection.

Statistical Analysis

All statistical analysis was performed in R version 3.5.0 using core distribution functions unless otherwise indicated¹⁷².

Results

Data Set Description

The work presented in this chapter was focussed on two complementary aspects of mutational processes in normal prostate epithelium. Firstly, mutational burden, corresponding mutational signatures and the presence of driver mutations was ascertained within histologically normal prostate tissues from a total of eight donors as listed in Table 1. Secondly, the clonal dynamics within individual glandular subunits could be explored for one of the eight donors as a whole prostate sample from this donor allowed detailed morphological reconstruction of these ductal structures.

To address both of these overarching topics, targeted microdissections of histologically normal prostate epithelium were obtained via LCM for all of these donors and subjected to WGS (Figure 39). For two donors, the WGS data was kindly provided by Luiza Moore. For the remaining six donors, tissue sections and corresponding WGS data was generated for this study. In total, 409 WGS samples were considered for the presented work. The majority of these samples were collected from the whole prostate sample of a 59-year-old donor and annotated according to their position within individual glandular subunits. These samples account for 319 WGS samples. For the remaining donors, two to thirty-two WGS samples were obtained and an overview is given in Table 1.

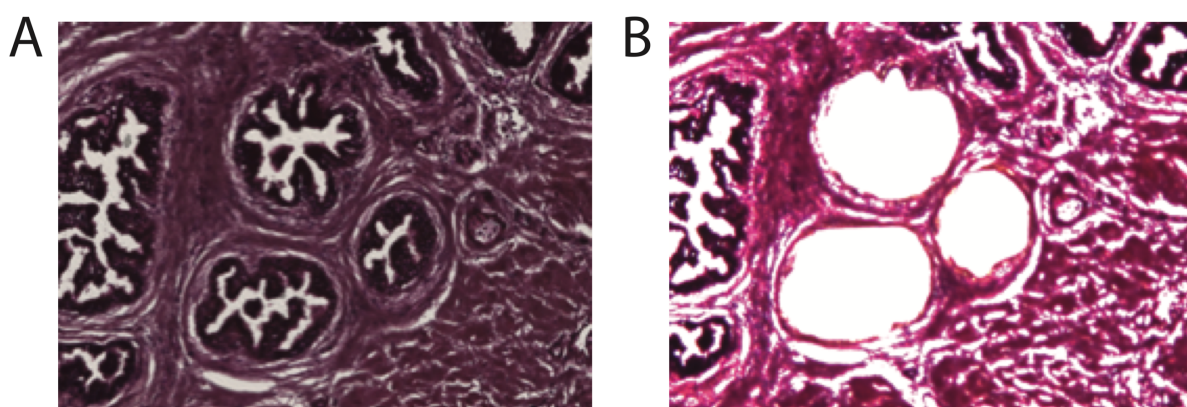


Figure 39: Example of Three Targeted Microdissections. Microdissections from histologically normal prostate epithelium were sampled using Laser-Capture-Microscopy. Microscopy images of the same area are displayed before **(A)** and **(B)** after collection of three distinct microdissections.

Table 1: Overview of Available Prostate and WGS Samples. WGS from microdissections of histologically normal prostatic epithelium was obtained from eight different donors. For six donors, WGS data was generated for the presented work and additional WGS data from two further donors was kindly provided by Luiza Moore.

Donor ID	Donor Age [Years]	Sample Type	WGS Samples
PD43390	22	Post-mortem biopsy	32
PD43391	31	Post-mortem biopsy	9
PD43392	47	Post-mortem biopsy	2
PD37885	59	WGS data	5
PD40870	59	Whole prostate	319
PD42298	70	Biopsy	21
PD43393	71	Post-mortem biopsy	9
PD28690	78	WGS data	12

Mutations Accumulate with Age in Normal Prostatic Epithelium

To consider a potential relationship between age and the accumulation of mutations in normal prostatic epithelium, base substitutions were called in the 409 WGS samples from LCM microdissections. While targeted microdissections aim to reduce cellular heterogeneity compared to traditional bulk sequencing and allow for calling of somatic variants in normal tissue, the sensitivity is still limited by the remaining degree of polyclonality and the obtained sequencing depth. The impact of polyclonality and sequencing depth on variant calling sensitivity in WGS data from LCM microdissections was estimated previously in our group by Simon Brunner. Since a robust estimate of polyclonality is essential for a correction of the observed mutational burden per microdissection, this correction could only be applied to 282 WGS samples. Notably, neither of the two WGS samples from the 47-year-old donor PD43392 could be considered in the final set of 282 WGS samples. Intuition and criteria for the sample selection and sensitivity correction are described in the methods section in more detail.

The sensitivity correction resulted in an average increase of 288 (SD = 241) mutations per microdissection. The average mutational burden observed per microdissection varied between donors with substantial intra-individual differences as listed in Table 2.

A negative-binomial GLM as described in the methods section was fitted to infer a potential age effect on the mutational burden within normal prostatic epithelium. Despite the considerable intra-individual variation, a significant age-effect across donors could be ascertained (Model Log-likelihood = -2125, fixed effect of $\log(\text{age}) = 0.83$, corresponding standard error = 0.13, $p < 3e-10$). After reversing the log-transformation that was necessary for the negative-binomial GLM and marginalisation over the random effect accounting for dependence of microdissections from the same donor, the best estimates from the model corresponded to an increase of 16 (SD = 0.3) mutations per year with a predicted offset of 150 (SD = 15) mutations (Figure 40).

Table 2: Corrected Mutational Burden in WGS Data from Microdissections of Normal Prostatic Epithelium. The mutational burden per microdissection was corrected for clonality and sequencing depth in appropriate WGS samples. While the average burden varies between donors, substantial intra-individual variation is also observed.

Donor ID	Donor Age [Years]	WGS Samples Considered	Mutational Burden (Mean \pm SD)
PD43390	22	25	439 \pm 175
PD43391	31	5	1096 \pm 329
PD43392	47	0	N/A
PD37885	59	3	1019 \pm 567
PD40870	59	208	991 \pm 589
PD42298	70	21	1371 \pm 498
PD43393	71	8	1255 \pm 849
PD28690	78	12	1530 \pm 160

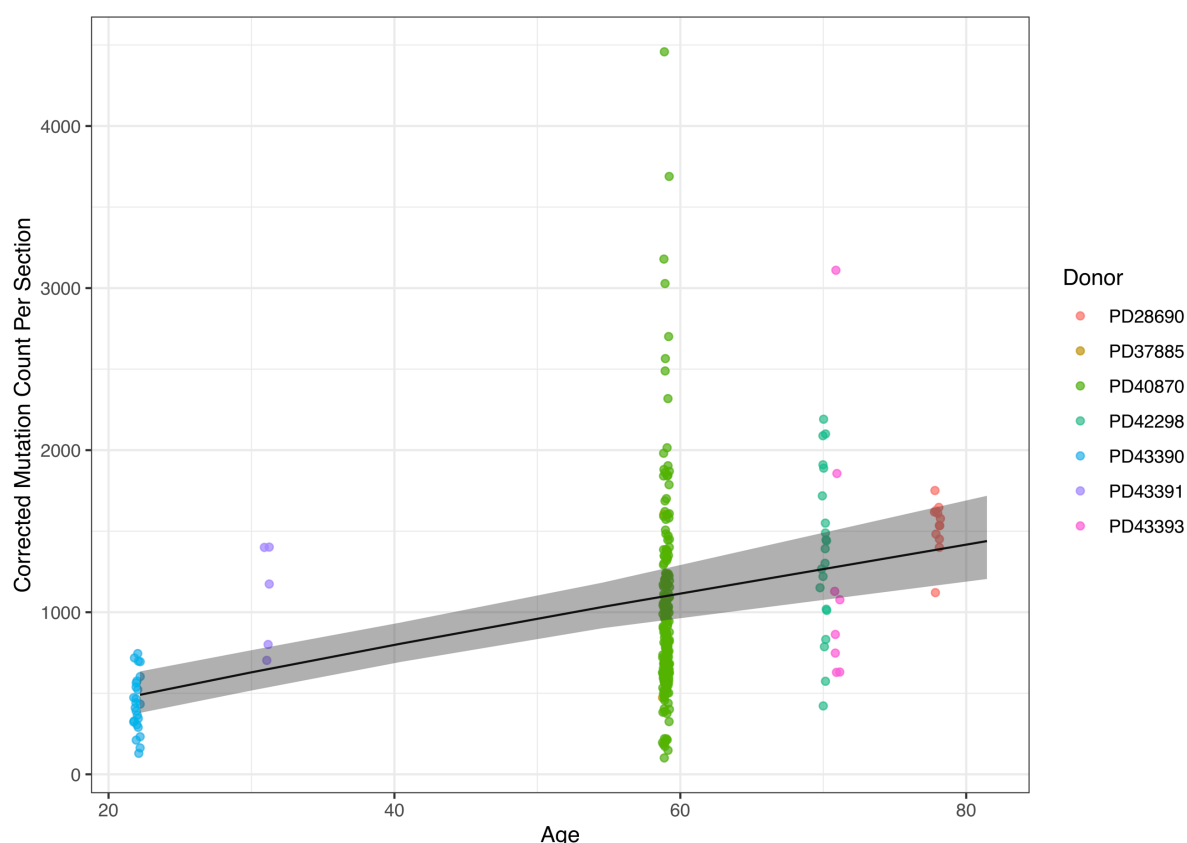


Figure 40: The Mutational Burden in Microdissections from Normal Prostatic Epithelium Increases with the Age of the Tissue Donor. Mutational burdens were ascertained in WGS samples and corrected for clonality and sequencing depth. A negative-binomial GLM revealed a significant age effect on the mutational burden per donor despite substantial intra-individual variation ($p < 3e-10$). The model suggests accumulation of 16 (SD = 0.3) mutations per year with an intercept at 150 (SD = 15) mutations. The best fit and corresponding 95% confidence interval are indicated in grey.

Clock-like Mutational Signatures Dominate the Mutational Landscape in Normal Prostatic Epithelium

Following the analysis of the total mutational burden in normal prostatic epithelium and its relationship with age, mutational signatures were extracted for all 409 WGS samples to reveal the underlying mutational processes. Across the eight donors, a total of five different mutational signatures could be observed. The main contribution to all microdissections from all eight donors were the clock-like mutational signatures SBS1 and SBS5, that were previously associated with cell division and the age of the individual^{36,66}. For the three younger donors between 22 and 47 years of age as well

as the 59-year-old donor PD37885 and the 70-year-old donor PD42298, these signatures accounted for 100% of the observed mutations with an average contribution of SBS1 of 17% (SD = 4%) and 83% (SD = 4%) of SBS5.

Similar average contributions of SBS1 and SBS5 could be observed in the microdissections in the 59-year-old donor PD40870 as well as in samples from the two oldest donors. However, in some of the microdissections from these donors, substantial presence of additional mutational signatures was detected. When additional mutational signatures could be detected, the relative contribution of SBS1 remained largely unchanged, while the contribution of SBS5 was decreased. The smoking signature SBS4 was detected in 3 of the 9 samples from 71-year-old donor PD43393, where they contributed to 25% (SD = 6%) of the mutational burden. Notably, the smoking status was unknown for all of the prostate donors and contribution of this smoking signature was only observed in a single patient. In seven microdissections from PD40870, SBS8 contributed to 22% (SD = 1%) of the observed mutational burden and SBS40 in 49 microdissections to 56% (SD = 12%). SBS40 was also observed in six of the 12 microdissections for the oldest donor PD28690, where they contributed to 52% (SD = 4%) of the mutational burden. While SBS8 and SBS40 do not have an associated aetiology, SBS40 has been shown to be correlated with patient's ages in some types of human cancer and thus, can display some clock-like features like SBS1 and SBS5 ³⁶.

In summary, the mutational landscape in normal prostatic epithelium was shown to be dominated by clock-like mutational signatures. While SBS1 and SBS5 accounted for 100% of the observed mutations in younger patients, additional signatures could be observed in microdissections from older patients with SBS40 being the only one identified in multiple patients (Figure 41).

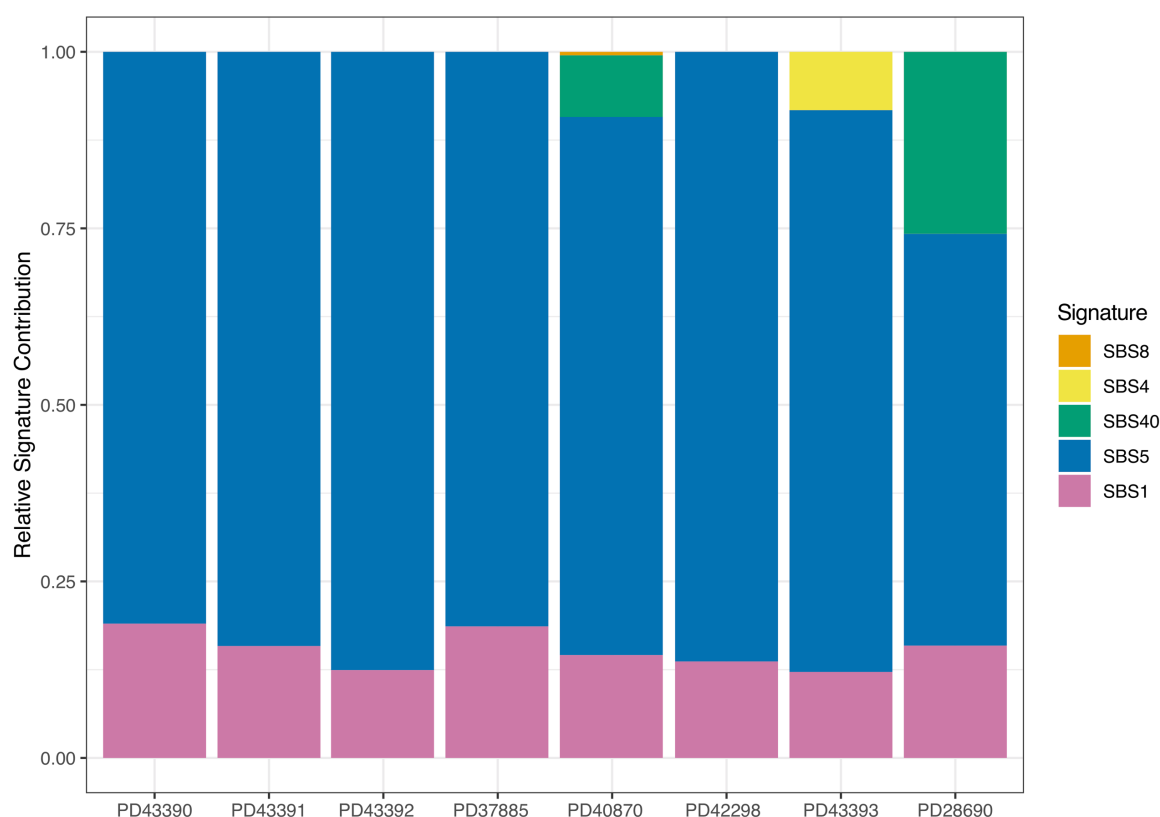


Figure 41: The Mutational Landscape in Normal Prostatic Epithelium is Primarily Shaped by Clock-like Mutational Signatures. For five of the donors, SBS1 and SBS5 contribute to 100% of the observed mutations. SBS4 and SBS8 are observed in a single patient with minor overall contributions while SBS40 is observed in increasing percentages in two older patients. SBS1, SBS5 and SBS40 all display clock-like features in several cancer types.

No Signs of Positive Selection and Low Frequency of Driver Mutations in Normal Prostatic Epithelium

Genes under positive selection feature an excess of non-synonymous mutations ¹¹⁷. To identify corresponding genes in normal prostatic epithelium, a dN/dS approach was applied to SNV calls from all 409 WGS samples that were available for this study as described in the methods section. After correction for multiple testing, no gene displayed significant signs of positive selection ($q > 0.1$ for all tested genes).

From the total amount of 153,706 unique mutations called in the 409 available WGS samples, 16 non-synonymous mutations were observed in 14 genes that were recently published as SMGs in prostate cancer (Table 3) ³¹⁷. From these sixteen mutations, one missense mutation in *RAG1* as well as one nonsense mutation in

SMARCA1 were identical to the ones observed in the discovery cohort that was used to define the catalogue of SMGs ³¹⁷. Furthermore, the established R219S driver mutation in *FOXA1* was observed in the 59-year-old donor PD40870 ³⁴¹⁻³⁴³. The remaining mutations did not overlap with previously annotated mutations or hotspots of mutations for prostate cancer as defined in the COSMIC Cancer Gene Census or the IntOGen database ^{127,339}. Thirteen of the non-synonymous mutations in SMGs were only detected in a single microdissection. The missense mutations in *COL15A1* and *PTPRC* were observed in two spatially adjacent microdissections each. Remarkably, R219S in *FOXA1* as the only established driver mutation detected in the presented data set, was detected in five spatially related microdissections.

Gene fusions involving the ETS transcription factor family are present in over half of all prostate cancers and are clinically used as marker for molecular subclassification ^{307,319}. Due to their high abundance in prostate cancer, the presence of relevant gene fusions was ascertained in microdissections in normal glandular epithelium. A total of 16 SV events as called by BRASS involved any of the ETS transcription factors or their most relevant fusion partners *TMPRSS2*, *SLC45A3* or *HNRP2AB1* across all 409 microdissections. Most of these events represented sequencing artefacts and no prostate cancer-relevant gene fusion could be detected.

Collectively, these results demonstrate that limited selection for base substitutions and a low amount of driver mutations existed in normal prostatic epithelium. However, the rare occurrence of driver mutations was implied to be associated with spatial expansion of the corresponding clone.

Table 3: Overview of Non-synonymous Mutations in Significantly Mutated Genes in Prostate Cancer. Sixteen mutations called in microdissections from normal prostatic epithelium were found in previously published SMGs in prostate cancer. Most mutations were only detected in a single microdissection. Notably, the known driver mutation R219S in FOXA1 was observed in five microdissections from the donor PD40870.

Donor ID	Gene	Change	Impact	Affected Samples
PD42298	<i>KMT2A</i>	C1479W	Missense	1
PD42298	<i>ZNF292</i>	Y2193C	Missense	1
PD40870	<i>RPRD2</i>	D327G	Missense	1
PD40870	<i>RAG1</i>	R34W	Missense	1
PD40870	<i>RAG1</i>	V529M	Missense	1
PD40870	<i>MRE11A</i>	I647L	Missense	1
PD40870	<i>BRCA2</i>	F1336L	Missense	1
PD40870	<i>FOXA1</i>	R219S	Missense	5
PD40870	<i>NCOR1</i>	P1665T	Missense	1
PD40870	<i>NOX3</i>	V72I	Missense	1
PD40870	<i>CHD7</i>	R1155C	Missense	1
PD40870	<i>COL15A1</i>	A542D	Missense	2
PD40870	<i>AR</i>	.	Essential Splice	1
PD40870	<i>AR</i>	R832Q	Missense	1
PD40870	<i>SMARCA1</i>	W426*	Nonsense	1
PD28690	<i>PTPRC</i>	E362K	Missense	2

Long-Ranging Glandular Subunits are Reconstructed from Serial Sections of a Whole Prostate Sample

A frozen whole prostate sample for the morphological reconstruction of individual glandular subunits was kindly provided by Rakesh Heer from Newcastle University. The prostate sample was received with a longitudinal cut through the AFMS, which exposed the prostatic urethra. After tissue fixation, a block of approximately 10 mm from the lower to mid prostate, which was centred around the verumontanum, was selected for the morphological reconstruction (Figure 42). Serial cross-sections alongside the prostatic urethra of 10 µm thickness were prepared by Yvette Hooks. Slide scans of the serial prostate whole-mount sections were manually screened and followed to reconstruct the complex branching pattern of individual glandular subunits. During the reconstruction process, samples were selected for subsequent microdissection to be able to overlay genomic with morphological information. The two glandular subunits that were reconstructed to the greatest detail comprised 91 and 88

WGS samples and could be traced across most of the distance from proximal to far peripheral parts of the prostate (Figure 43). These glandular subunits were selected from the left- and the right-hand side of the prostate and covered a cubic area of 33.4 mm^3 and 32.7 mm^3 , respectively (Figure 44).

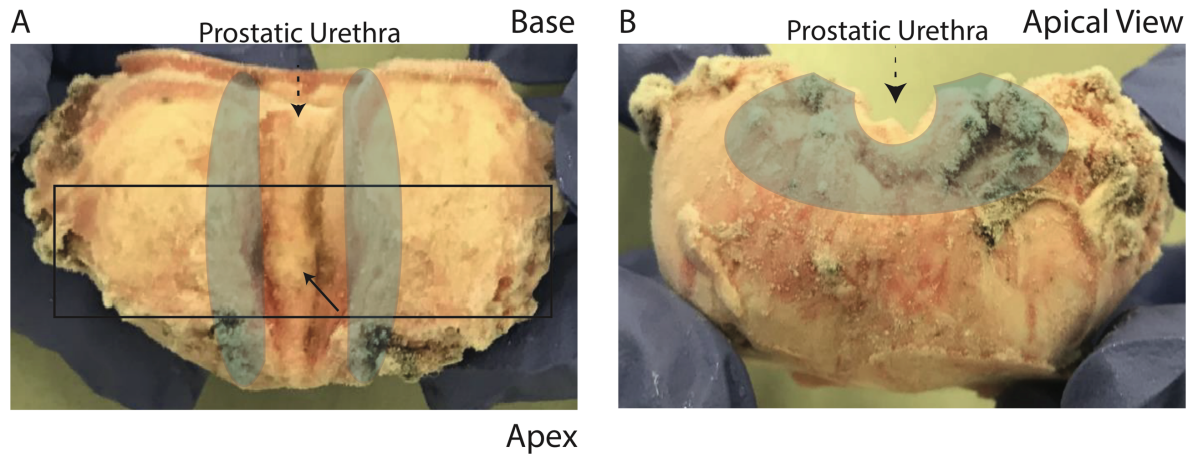


Figure 42: Overview of Unprocessed Whole Prostate Sample. A whole prostate sample from a 59-year-old-donor was used to explore the clonal dynamics within individual glandular subunits. The prostatic urethra was exposed through a longitudinal cut alongside the AFMS. A block of approximately 10 mm (indicated by a black box) around the verumontanum (indicated by the straight-line black arrow) was selected for serial cross-sectioning alongside the prostatic urethra. Glandular subunits originate in the proximal prostate (indicated in grey shading) where they drain into the urethra and extend into the periphery in a complex branching pattern. **(A)** Overview of the whole prostate sample. **(B)** Apical view to illustrate direction of sectioning.

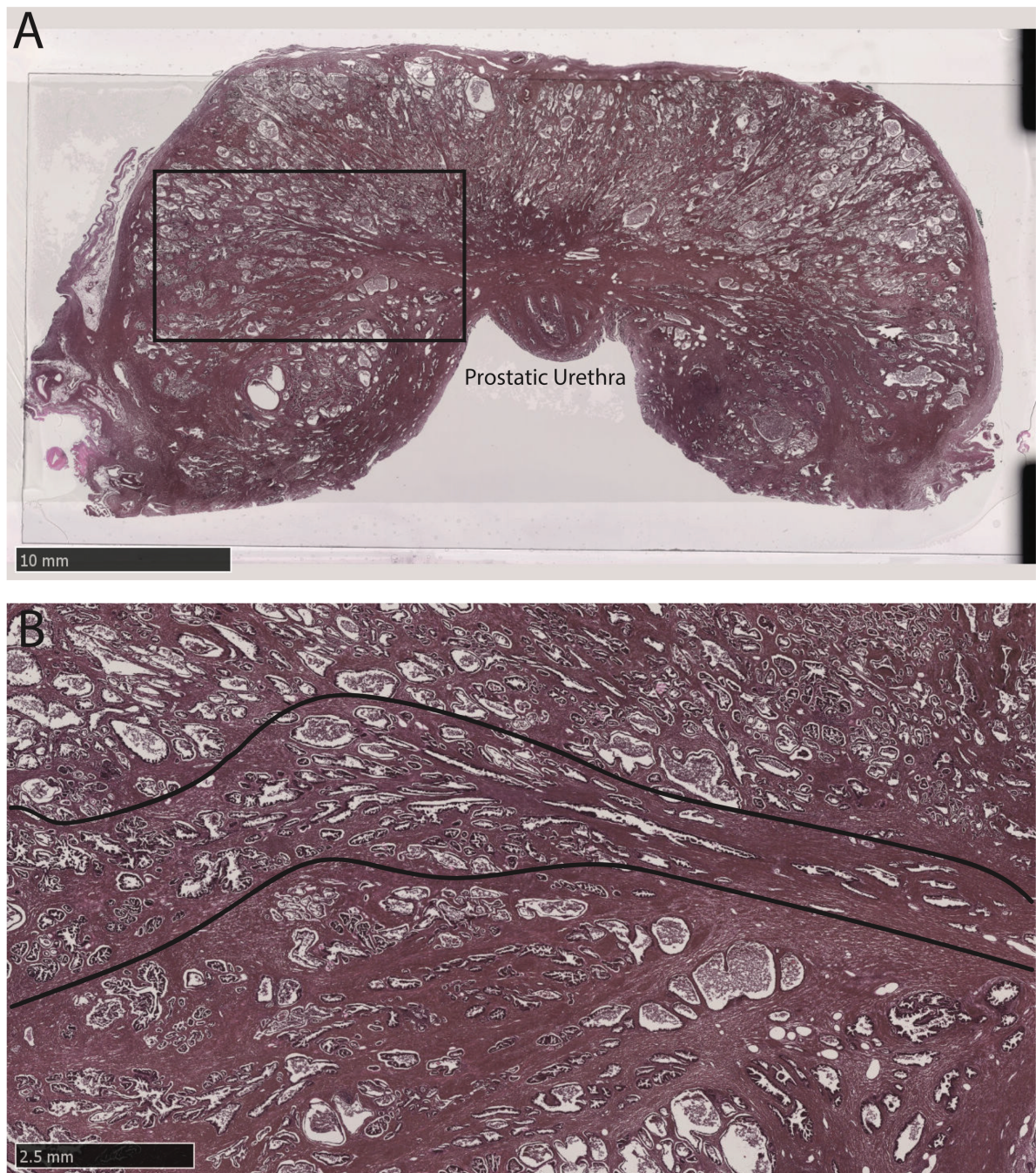


Figure 43: Exemplary Cross-section and Part of Glandular Subunit. Serial cross-sections were obtained from a whole prostate sample and manually screened for long-ranging glandular subunits that extend from the urethra-proximal to peripheral areas of the prostate. **(A)** H&E-stained cross-section of the whole prostate sample. Large amounts of ductal and acinar tissue can be seen enclosing cell-free brighter areas within the uniformly stained stromal tissue. The area indicated by the black box contains a long-ranging glandular subunit and is shown in more detail in **(B)**. **(B)** Part of a long-ranging glandular subunit is outlined in black. Following the epithelial structures in the adjacent serial sections, connections between the highlighted glandular tissue could be observed and were used to reconstruct the branching pattern and extent of individual glandular subunits.

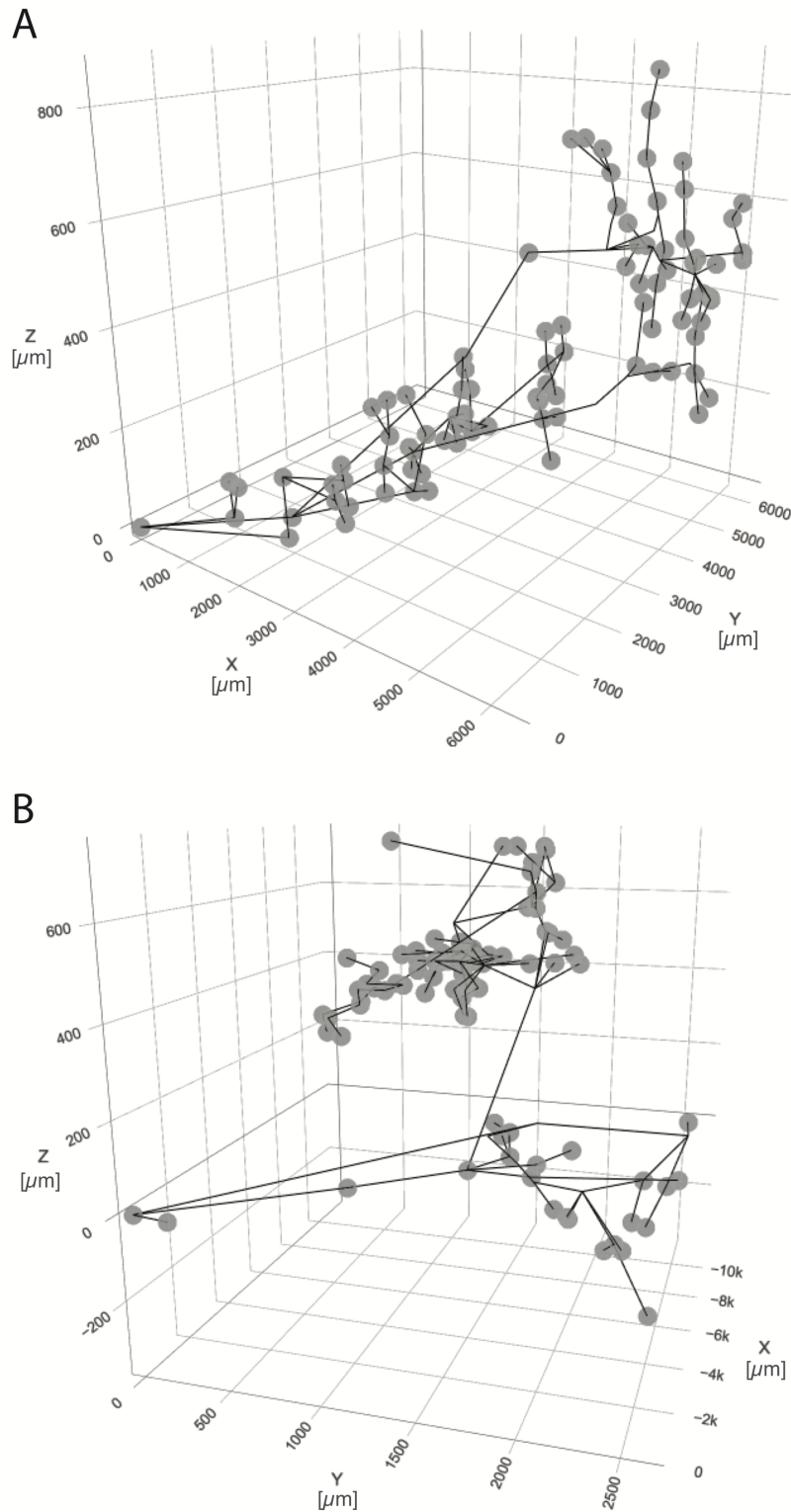


Figure 44: Overview of 3D-Reconstruction of Glandular Subunits. Individual glandular subunits were manually traced and reconstructed from serial cross-sections of a whole prostate sample. Grey globes indicate a microdissection sample with WGS data and the connecting black lines indicate the branching pattern of the ductal network. The microdissection sample at the origin is the most proximal sample of the respective glandular subunit. **(A)** Glandular subunit from the left-hand side of the prostate with 91 microdissection samples. **(B)** Glandular subunit from the right-hand side of the prostate with 88 microdissection samples.

Increased Rates of Mutation Accumulation and Cell Division are Observed in the Peripheral Parts of Individual Glandular Subunits

Substantial intra-individual variation in the mutational burden of microdissections could be observed during the analysis of age-dependent accumulation of mutations (Figure 40). Notably, a clock-like pattern of mutation acquisition was also supported by the mutational signatures that are active in normal prostatic epithelium (Figure 45). Due to the morphological reconstruction of glandular subunits that was possible for donor PD40870, a spatial effect on the mutational burden within one donor could be considered. Such an effect could be confirmed as the mutational burden was significantly increasing in a proximal to distal fashion within individual glandular subunits (0.06 mutations per μm , $p < 8\text{e-}6$, OLS; Figure 45A). Moreover, the mean telomere length per microdissection decreased in the same direction (-0.08 bp per μm , $p = 1\text{e-}5$, OLS) and telomere length was shown to be inversely correlated with the mutational burden in microdissections from normal prostatic epithelium (-0.32 bp per mutation, $p < 1\text{e-}10$, OLS; Figure 45B). Notably, the observed clonality per microdissection was not correlated with the ductal distance of the sample, suggesting the change in mutational burden and telomere length did not depend on the clonal structure but rather were independent and local effects ($p = 0.7$, OLS; Figure 45C). Collectively, these results demonstrate that mutation accumulation and cell division are tightly interlinked in normal prostatic epithelium. Furthermore, rates of cell division and corresponding accumulation of mutations are variable within the ductal networks and significantly increase in a proximal to distal or ductal to acinar direction.

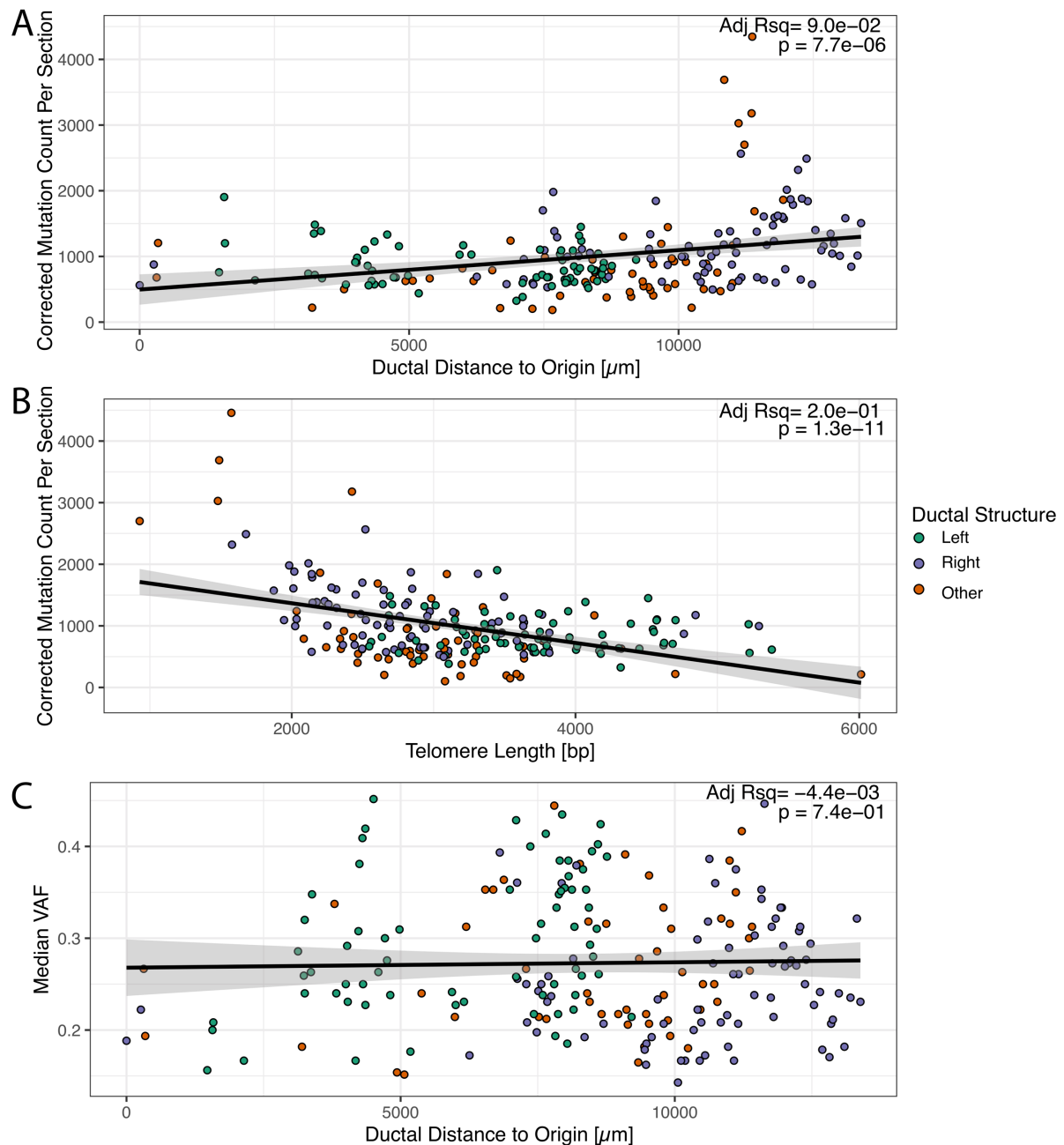


Figure 45: Mutational Burden and Cell Division Increase in Proximal to Distal Fashion within Glandular Subunits. Data points are coloured to distinguish between the two main glandular subunits discussed in this chapter and additional ones not extensively discussed here. **(A)** The mutational burden per microdissection increases with the ductal distance to the most proximal sample per glandular subunit. **(B)** Telomere length and mutational burden are highly correlated in prostatic ductal epithelium, pointing out cell division and parallel processes as main drivers for mutation accumulation. **(C)** The median VAF per microdissection as a measure for clonality does not change with ductal distance to origin. The increased mutational burden and decreased telomere length along the proximal to distal axis is not an effect of the clonal structure of the microdissection.

Complex Lineage Relationships are Observed Within Glandular Subunits

Information about somatic variants can be used to retrospectively infer the lineage relationship of somatic cells. Under the assumption of the infinite site model of evolution, shared mutations between currently existing cells imply the existence of a MRCA for all cells carrying this mutation³⁹. In case of single cell or single cell-derived organoid data, the lineage tree can in theory directly be inferred from a matrix of mutation identity across all considered samples^{82,86,344}. However, microdissections from prostatic epithelium comprised several hundred to a few thousand somatic cells, which made consideration of the clonal structure within one microdissection necessary before generation of a lineage tree. Microdissections from the left- and right-hand side of the prostate reconstructed to most detail displayed median VAFs of 27% (SD = 14%) and 26% (SD = 13%), respectively. These moderate median VAFs suggested that cells within one microdissection did not share a very recent MRCA. While glandular epithelium was targeted in the microdissections, residual contamination with surrounding stromal cells would explain a distant MRCA. Notably, detachment of the epithelial structures from the surrounding stroma was frequently observed as common artefact from tissue processing. These structures could be sampled with minimised risk for stromal contamination and displayed a similar degree of polyclonality. Therefore, the observed median VAFs were more likely to be linked to the existence of several epithelial clones within one microdissection and the degree of polyclonality rendered lineage tree inference strictly based on mutational identity inappropriate.

While lineage relationships between microdissections could not directly be inferred from a mutation identity matrix, corresponding lineage trees could be derived by identifying ancestral clones based on discrete clusters of mutations at similar VAF and the relative contribution of the ancestral clones to individual microdissections. This is explained in the methods section in more detail. For the left- and right-hand side structure with 91 and 88 WGS samples, a total of 25,163 and 62,759 mutations were grouped into 64 and 79 clusters, respectively. While most clusters only contributed to a limited number of microdissections, for both glandular subunits three clusters could be identified that contributed evenly to a low percentage in all microdissections. Upon manual inspection, these clusters were shown to nearly exclusively include mutations in adjacency to short mononucleotide stretches or from repetitive genomic regions

(Figure 46). These artefactual clusters as well as all clusters with less than ten mutations were excluded before manual inference of the lineage tree based on the pigeonhole-principle. From the remaining 39 and 49 clusters for the left- and right-hand structure, 37 and 43 clusters could be placed on the respective lineage trees for the two glandular subunits accounting for 19,624 and 31,046 mutations. The clusters that could not be placed during this step represent a resolution limitation of the clustering algorithm used. Especially mutations at low VAFs in one or few samples cannot correctly be placed. Therefore, the presented phylogenies are an approximation with greater inaccuracy in terminal branches.

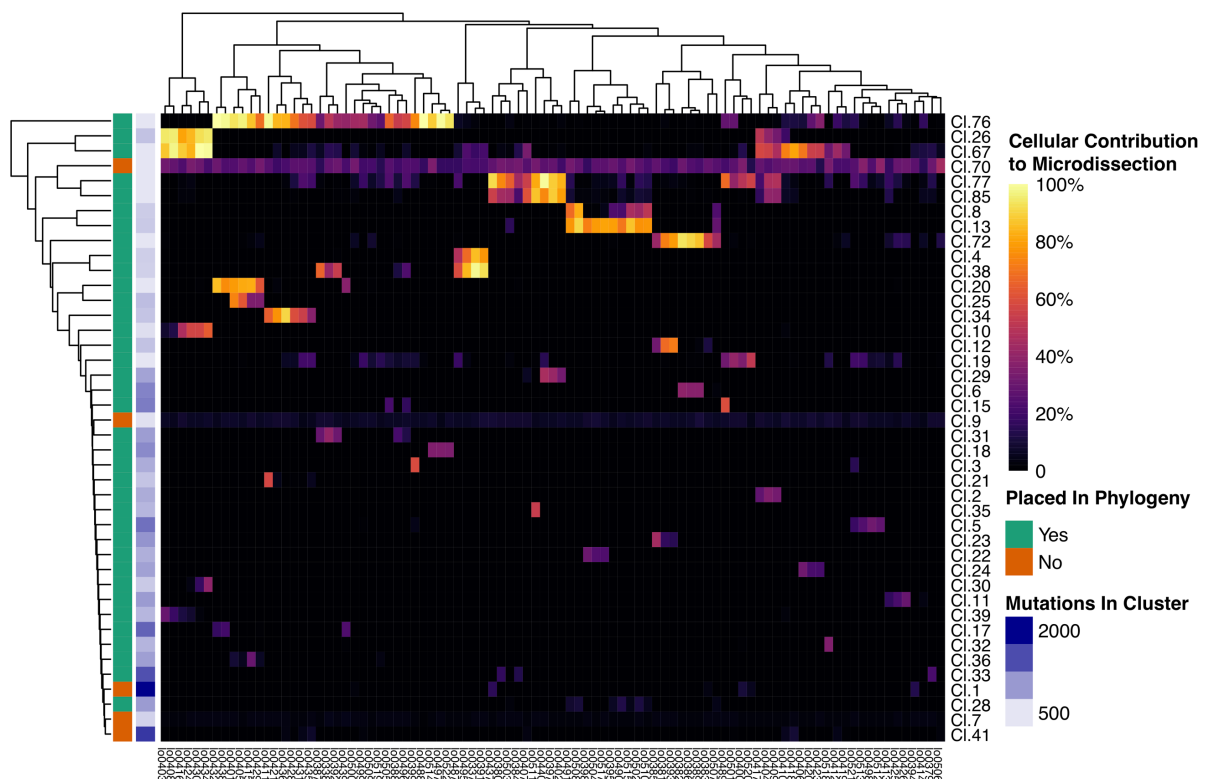


Figure 46: Discrete Clusters of Mutations Identify Ancestral Clones that Contribute to Current Prostatic Epithelium. Clusters of mutations are identified based on VAF using an n-HDP clustering algorithm and constitute the rows of the heatmap. These clusters of mutations represent ancestral clones which contribute to current cells sampled in individual microdissections, which are shown in the columns of the heatmap. The median VAF of cells associated with an ancestral cluster is used to estimate the cellular contribution to each microdissection and is visualised using colour intensities. The co-occurrence of multiple ancestral clones within the same microdissection sample and their corresponding cellular contribution was used to construct the lineage tree for individual glandular subunits. The shown example corresponds to the glandular subunit on the left-hand side of the prostate. Cluster 70, 9 and 7 include nearly exclusively artefactual mutations and were excluded from the lineage tree. Furthermore, cluster 1 and 41 are insufficiently split by the clustering algorithm used and could not be placed unambiguously.

The two inferred phylogenetic trees revealed complex lineage relationships within individual glandular subunits and shared common patterns between both structures. The 37 clusters from the left-hand side glandular structure were divided into nine distinct clades (Figure 47). Four of these clades included 28 of the 37 clusters and featured several parallel as well as nested branches. These four clades were all defined by an initial cluster of only 13 – 16 mutations, named Cluster 67, 72, 76 and 77 respectively, according to the random naming convention of the n-HDP clustering algorithm. Notably, the most terminal branches in these clades extended to approximately 1,000 - 1,500 mutations. Data presented earlier showed that somatic mutations accumulate linearly with age (Figure 40), and unpublished work from our laboratory has indicated that fetal tissues at around 18 weeks of gestation have accumulated 20 – 40 mutations per cell (MS Chapman, Personal Communication). Therefore, these four clusters with 13 – 16 mutations each must represent ancestral clones that existed very early in life, likely during embryonic development. Directly underneath Cluster 76 and Cluster 77, three short branches of 18-21 mutations are observed defined by Cluster 20 as well as Cluster 19 and 85. After these early coalescences that also potentially occurred during embryonic development, a second group of seven coalescence events was observed in the clades defined by Cluster 67 and 76 as well as two further clades defined by Cluster 13 and 38. Considering the total branch length of these coalescence events, they occurred after the accumulation of 414 (SD = 117, range 248 - 602) mutations on average. According to the age regression, the average mutational burden of 414 is reached after 16 to 17 years. Remarkably, the observed mutational burden for these clones is similar to, but slightly lower than the mutational burden observed in microdissections from the 22-year-old donor PD43390 with average corrected burden of 439 (SD = 175, range 129 - 745) mutations. Therefore, this second wave of coalescences is likely to have occurred during adolescence. While coalescence events were observed during embryonic and pubertal development in six of the nine clades, three additional polytomous clades without further sub-nesting and respective burden of 527, 795 and 1,218 mutations were observed (Figure 47).

For the right-hand side glandular subunit, a lineage tree with 16 clades comprising 43 ancestral clones could be inferred (Figure 48). Again, four clades were defined by initial coalescences after 10 – 27 mutations and included 25 ancestral clones in

parallel and sub-nested branches. The ancestral clones corresponding to Cluster 8, 61, 71 and 98 were again likely to represent embryonic development. A second group of coalescence events was observed in three of these clades as well as in four further clades defined by Cluster 11, 16, 31 and 119. Similar to the second group of coalescence events observed in the lineage tree for the previously described glandular subunit, these coalescence events were associated with an average mutational burden of 338 (SD = 66, range = 193 - 415) mutations. Therefore, these events were likely to have happened during pubertal development. In addition to the eight clades with embryonic or pubertal coalescence events, a total of eight additional clades without any internal coalescence events was observed (Figure 48).

While no further coalescence event was observed in the lineage tree for the glandular subunit on the left-hand side of the prostate after the events associated with pubertal development, further sub-nesting in the lineage tree of right-hand side structure could be observed in the clade defined by Cluster 98 (Figure 48). Here, Cluster 118 was identified as an ancestral clone that existed with a total mutational burden of 981 mutations. Moreover, Cluster 36 was identified within the progeny of Cluster 118 with an additional 267 mutations. These coalescence events were unique and could not be dated back to embryonic or pubertal development. Thus, they potentially happened during adult tissue homeostasis. Remarkably, the R219S driver mutation in the *FOXA1* gene identified in a previous analysis was associated with Cluster 118 (Table 3).

In summary, the analysis of lineage trees from two glandular subunits of the same donor revealed complex lineage relationships in the long-ranging ductal network. Coalescence events were likely to be associated with the time periods of embryonic and pubertal development. Moreover, the only observed coalescence events that could reliably be associated with the time period of adult tissue homeostasis were linked to the presence of a prostate cancer driver mutation.

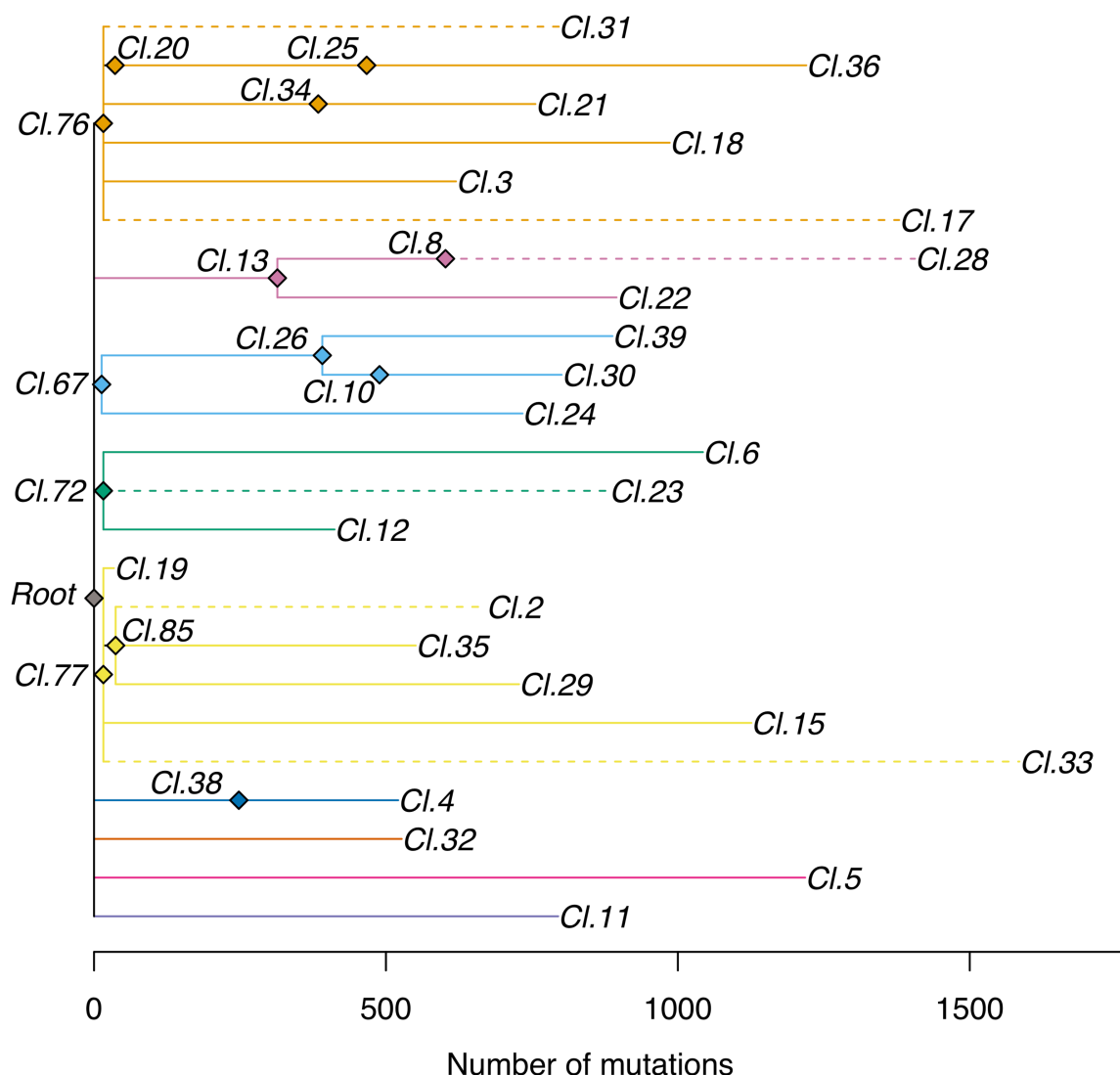


Figure 47: Phylogenetic Tree of Prostatic Epithelium From a Glandular Subunit of the Left-Hand Side of the Prostate. Mutations called in microdissections from prostatic epithelium were clustered to reveal the presence of ancestral clones that gave rise to the sampled cells. These ancestral clones were arranged in a lineage tree based on their co-occurrence pattern and cellular contribution to individual microdissections. Solid lines indicate nesting with absolute confidence, while relationships marked by dashed lines are highly supported but not strictly necessary (see methods section). Four main clades defined by ancestral clones 67, 72, 76 and 77 comprised 28 of the 37 identified clones in this lineage trees. These ancestral clones as well as Cluster 19, 20 and 85 were dated to embryonic development. A second group of coalescence events was observed at average mutational burden of 414 mutations. These ancestral clones were likely to have existed during pubertal development.

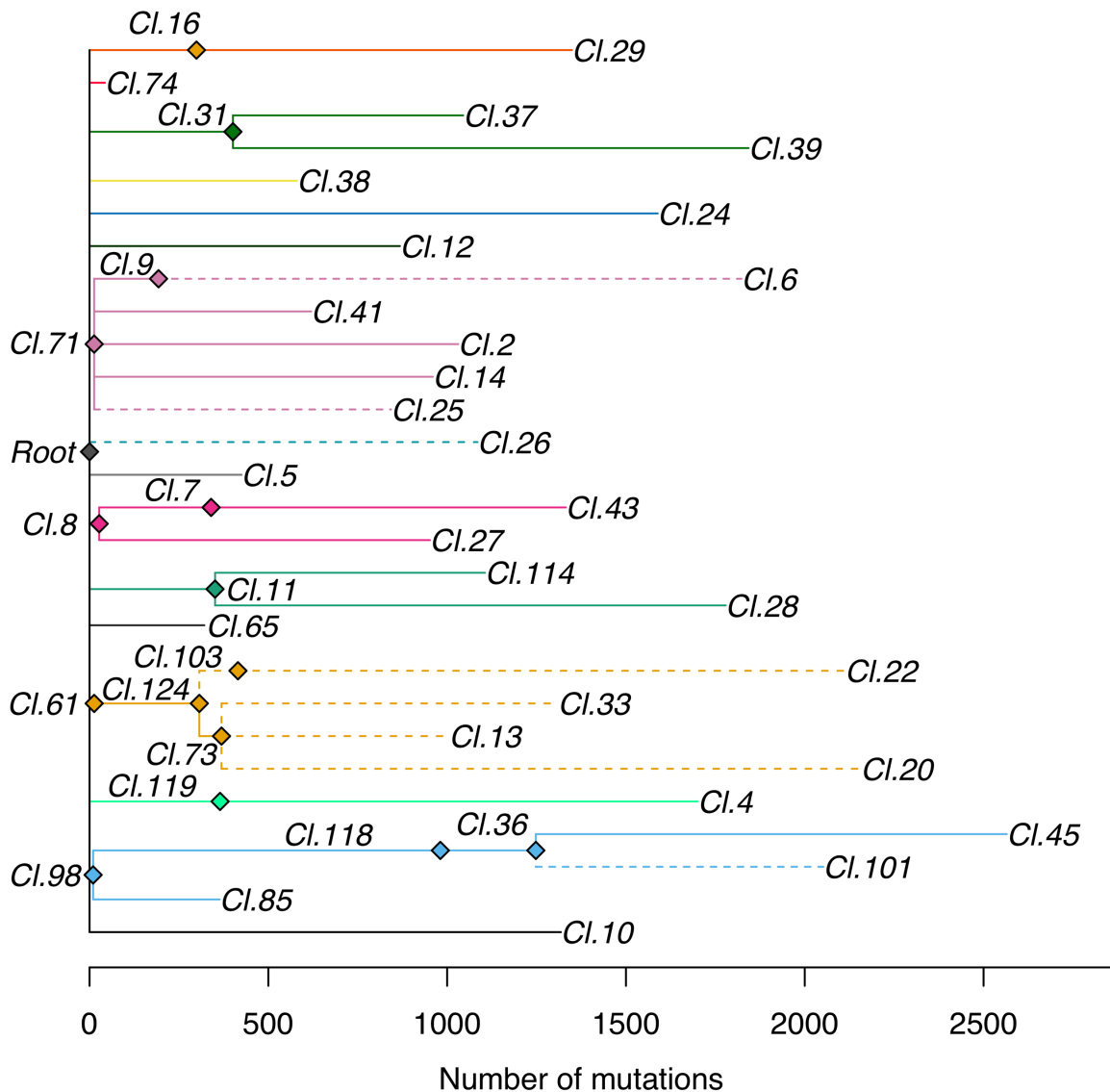


Figure 48: Phylogenetic Tree of Prostatic Epithelium From a Glandular Subunit of the Right-Hand Side of the Prostate. Mutations called in microdissections from prostatic epithelium were clustered to reveal the presence of ancestral clones that gave rise to the sampled cells. These ancestral clones were arranged in a lineage tree based on their co-occurrence pattern and cellular contribution to individual microdissections. Solid lines indicate nesting with absolute confidence, while relationships marked by dashed lines are highly supported but not strictly necessary (see methods section). The four clades defined by Cluster 8, 61, 71 and 98 revealed ancestral clones that were likely to have existed during embryonic development. The majority of further coalescence events was observed at total mutational burdens of 193 – 415 mutations, which point to ancestral clones that have existed during adolescence. Notably, Cluster 118 and 36 were identified as unique ancestral clones that must have existed during adult tissue homeostasis. Cluster 118 included an established prostate cancer driver mutation in *FOXA1*.

Directed Morphogenesis is Observed During Embryonic and Pubertal Development of Glandular Subunits

The relative cellular contribution of ancestral clones to microdissections from two reconstructed glandular subunits was explored in their corresponding 3D-models to consider the interdependence of phylogenetic and morphological relationships (see Figure 44 for an overview of the 3D models). In general, similar observations could be made in both glandular subunits.

Clusters that were dated back to embryonic development displayed a very wide and mostly contiguous distribution across several different main and side ducts. For the left-hand glandular structure, where a higher percentage of all clones was found to be nested within one of the four main clades defined by such an embryonic development clone, most microdissections displayed a contribution of greater than 70% of one of these four clones. In several examples, the relative cellular contribution of each of these clones to individual microdissections increased in a proximal to distal direction. Remarkably, while their general spatial distribution largely overlapped, the contribution to individual microdissections was largely mutually exclusive (Figure 49).

More recent clones still displayed a contiguous but more spatially confined distribution. Clones that were nested in the phylogenetic tree also showed a matching nesting in the anatomical ductal tree - that is, the subclone of an ancestral clone was confined to a subsection of the duct encompassed by the ancestral clone. Parallel branches within the same clade usually displayed a mutually exclusive distribution. While clones that existed during embryonic development were found to be widely distributed across several main and side-branches of the ductal network, the clones associated with pubertal development were mainly found in one or few directly adjacent side-branches (Figure 50). Cellular contribution of pubertal clones usually increased in proximal to distal direction and commonly reached over 50%. Since even proportions of luminal and basal cells constitute normal prostatic glandular epithelium, MRCA's with cellular contribution greater than 50% are at least bipotent ²⁹⁶.

Collectively, the spatial distribution of ancestral clones that were associated with embryonic and pubertal phases supported a model of directed morphogenesis during prostate development. Remarkably, only few clones were implied to initially give rise to the ductal network during embryonic development and another more spatially confined expansion in proximal to distal direction was suggested during adolescence.

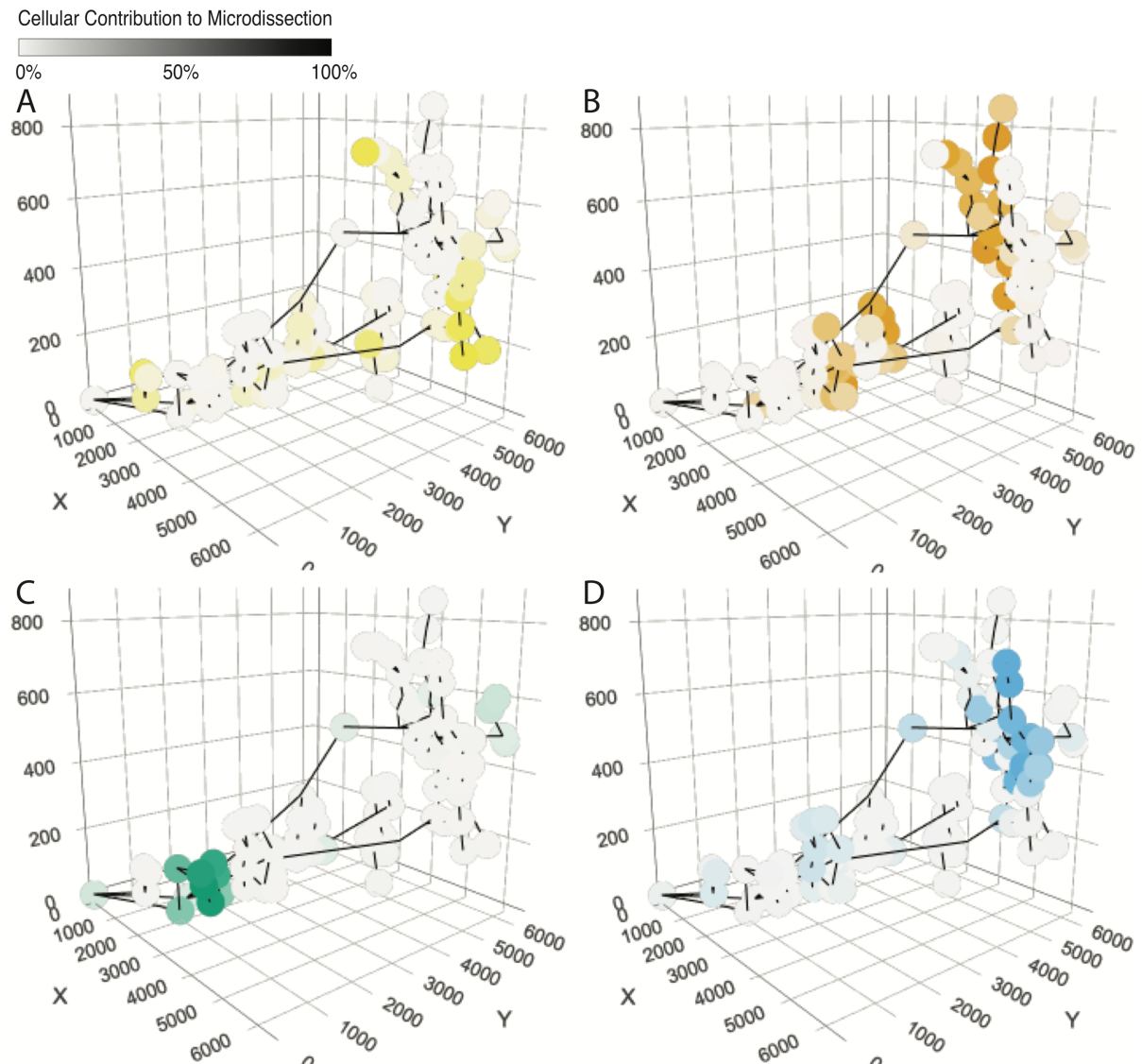


Figure 49: Few Ancestral Clones Contribute to a Whole Glandular Subunit During Embryonic Development. The cellular contribution of four main ancestral clones that were dated back to embryonic development during lineage tree reconstruction of the glandular subunit from the left-hand side of the prostate is displayed. Light grey colours indicate absence of clones while more intense colours show higher contributions to the cell mass comprising the corresponding microdissection. These four clones display a wide and mainly contiguous spatial distribution. While their general distribution overlaps, the contribution to individual microdissections is largely mutually exclusive. **(A)** Cluster 77. **(B)** Cluster 76. **(C)** Cluster 72. **(D)** Cluster 67.

Cellular Contribution to Microdissection
 0% 50% 100%

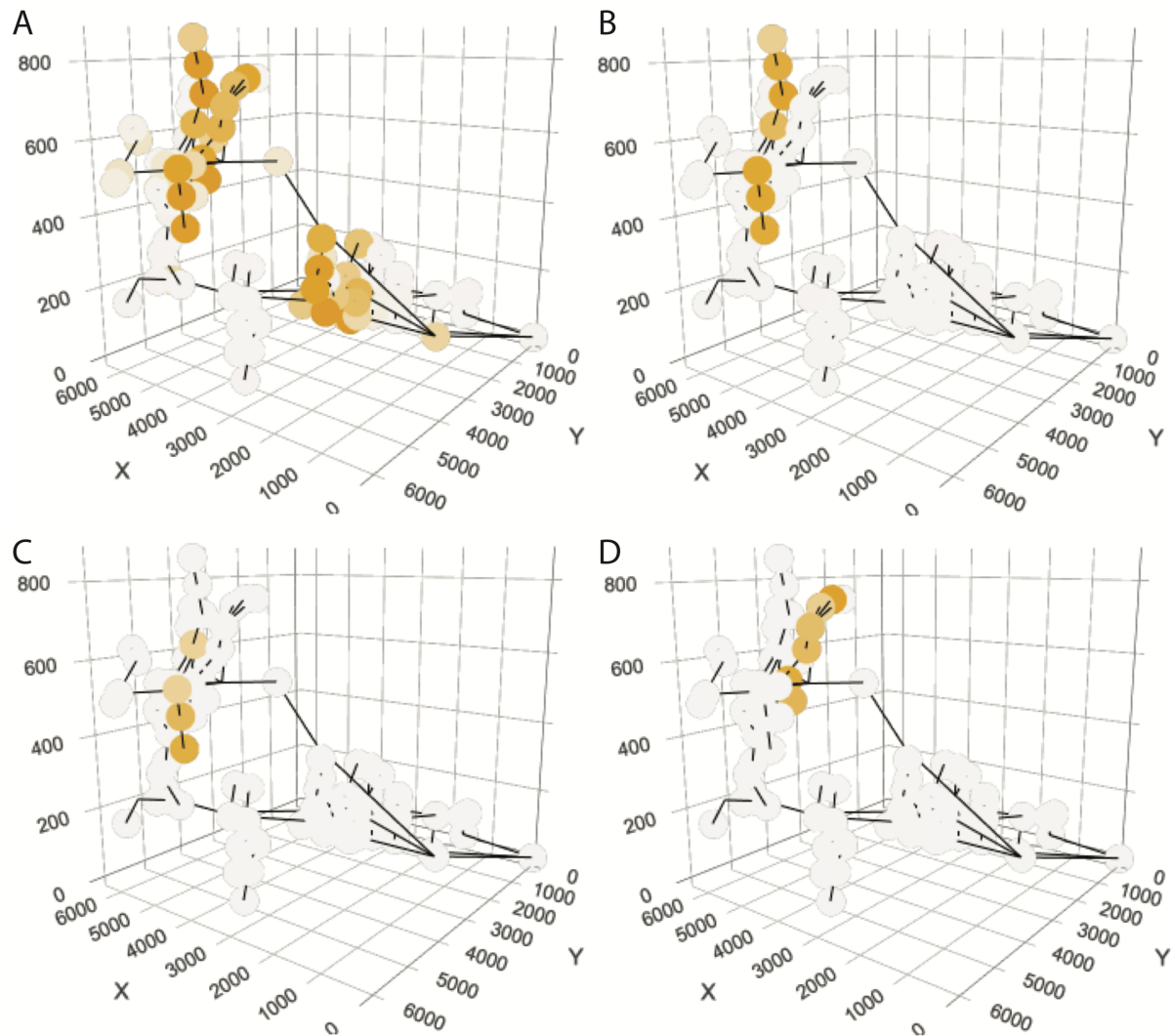


Figure 50: Phylogenetic Relationships are Recapitulated by the Spatial Distribution of Clones. Clones that were nested underneath more ancestral clones in the lineage tree displayed a contiguous but more spatially confined distribution. Parallel branches usually displayed a mutually exclusive spatial distribution. **(A)** Cluster 76 represented a clone that was dated back to embryonic development within the glandular subunit on the left-hand side of the prostate. **(B)** Cluster 20 was nested underneath Cluster 76 in the lineage tree and showed a more spatially confined distribution within the area covered by Cluster 76. **(C)** Cluster 25 was further nested underneath Cluster 20 on the lineage tree, which was also reflected by an even more restricted spatial distribution. Cluster 25 was dated to have existed during adolescence. **(D)** Cluster 34 represented another ancestral clone during adolescence and was nested underneath Cluster 76. Notably, it represents a parallel branch to Cluster 20 and 25. Cluster 34 covered a contiguous area within Cluster 76 but did not display a spatial overlap with the clones defined by Cluster 20 and 25.

Local Proliferation of Progenitor Cells is Suggested During Adult Tissue Maintenance

Similar to the trend that pubertal clones were observed with more confined spatial distribution compared to embryonic clones, the more recent terminal nodes generally displayed an even more confined but still contiguous distribution. Terminal nodes of clusters with hundreds of mutations were commonly only observed in few microdissections. Similar to the observation for ancestral and pubertal clones, the cellular contribution to microdissections often increased in proximal to distal fashion.

However, few examples with the highest contribution to more proximal microdissections could be observed. For example, Cluster 36 that was nested underneath the embryonic clone Cluster 76 and 25 as well as the pubertal clone Cluster 20 was observed in three contiguous microdissections from a peripheral side duct (Figure 47 and 51A). Within these microdissections, Cluster 36 contributed to 29% of the most proximal sample, 5% of the middle one and 10% of the most distal sample. Cluster 36 was associated with 753 mutations and only 176 and 230 of these mutations were detected in the two more peripheral samples. The sequencing depth at the remaining 523 – 577 loci ranged between 21 – 28 in the peripheral samples, setting the VAF detection limit at about 4%. Notably, the median VAF of the detected mutations was substantially higher with 9% and 13% and potentially suggested an incomplete split of Cluster 36 during the n-HDP-based identification of ancestral clones.

Notably, the most proximal sample with the highest contribution of Cluster 36 featured the lowest clonality of the three microdissections with a global median VAF of 24%. In contrast, the two more peripheral microdissections featured median VAFs of 37% and 33%, respectively. The higher clonality in the more distal samples suggested that absence of the remaining 523 – 577 mutations associated with Cluster 36 in the more peripheral microdissections was not due to a technical limitation of the WGS approach. Remarkably, the distance between the most proximal and middle microdissection of the presently described side duct were only 140 μm (Figure 51A).

Several other examples for substantial local accumulation of mutations could be observed like Cluster 21 for the glandular subunit on the left-hand side that marked

the acquisition of at least 300 mutations across a distance of just 250 μm . Notably, Cluster 21 contributed to 58% of the cells in one microdissection (Figure 51B). Contributions to over 50% of the cellular mass of microdissections from glandular epithelium were observed for several terminal branches that marked MRCAs, which have existed during the time period of adult tissue maintenance. As explained for the pubertal clones, these MRCAs must have had at least bipotent lineage potential as they had to give rise to luminal as well as to basal cells.

In summary, most terminal nodes that marked the period of adult tissue homeostasis were strongly spatially confined, commonly displayed at least bipotent lineage potential and acquired hundreds of private mutations across a distance of a few hundred micrometres. Hence, the presented data suggested extensive local proliferation of progenitor cells for adult tissue maintenance.

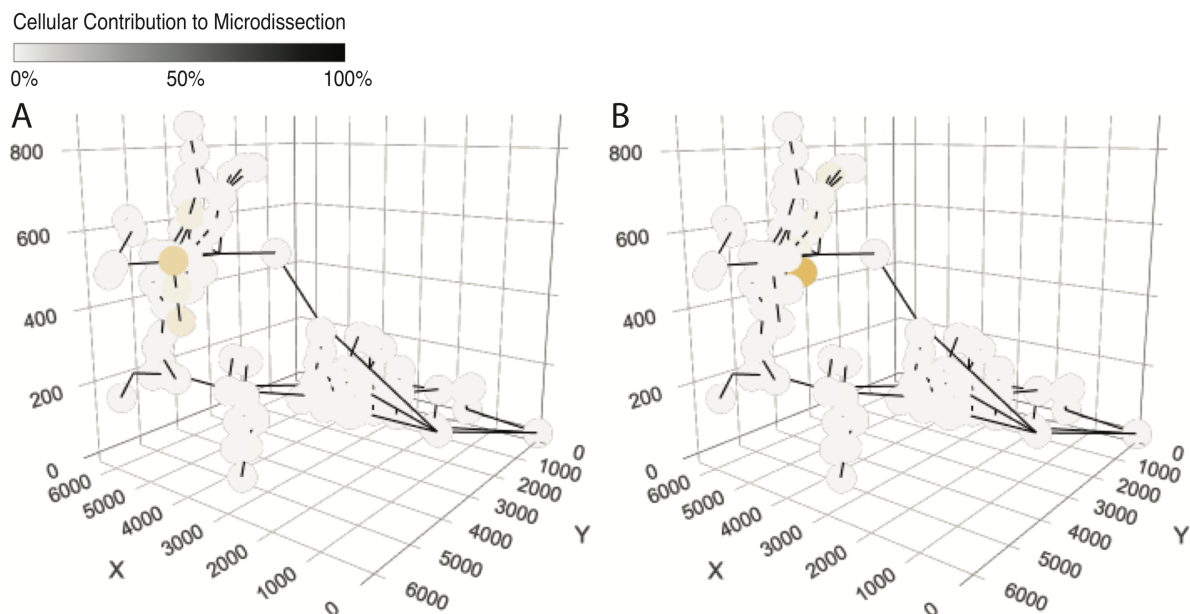


Figure 51: MRCAs That Existed During Adult Tissue Homeostasis Are Strongly Spatially Confined and Display High Amounts of Private Mutations. (A) Cluster 36 from the left-hand side glandular subunit is associated with 753 mutations. Substantial evidence for the presence of Cluster 36 is only observed in the most proximal microdissection from a peripheral side duct. A total of three microdissections were sampled from this duct but no support for over 500 of Cluster 36-associated mutations is found in the two more peripheral samples. The ductal distance between these microdissections was 140 μm . **(B)** Cluster 21 from the left-hand side glandular subunit contributes to over 50% of the sampled cells in a single peripheral microdissection. No evidence for over 300 mutations associated with Cluster 21 could be observed in the adjacent microdissection despite similar clonality in these samples. The ductal distance between these microdissections was 250 μm .

A Prostate Cancer Driver Mutation is Associated with Spatial Expansion of an Adult Clone

With the exception of all but two observed coalescence events, ancestral clones that were widely distributed across the ductal network, were likely to be linked to embryonic or pubertal development. However, Cluster 118 and Cluster 36, which was nested underneath Cluster 118, were found in the glandular subunit of the right-hand side and represented ancestral clones with a total burden of 981 and 1,248 mutations, respectively. The clade defined by Cluster 118 was nested underneath the embryonic clone Cluster 98 and the pubertal clone Cluster 85 was found in parallel (Figure 48).

In contrast to the limited spatial distribution of other adult clones across few microdissections within one or few spatially related ducts, cellular contribution of Cluster 118 was prominent in 13 microdissections covering one main peripheral duct with four outgoing branches. Cellular contribution of Cluster 118 ranged between 40 – 80% in these microdissections. Cluster 36 was present in most of these microdissections as well but its cellular contribution only exceeded 20% in 6 microdissections in two spatially related side branches. Cluster 118 displayed greater spatial distribution than the parallel pubertal clone Cluster 85 and covered most of the spatial territory of its embryonic clone Cluster 98 (Figure 52).

As explained during description of the lineage trees, Cluster 118 contained a prostate cancer driver mutation in *FOXA1*. Considering the distance to the closest followed branching points before substantial contribution of Cluster 118 was detected, a total of about 4 mm of spatially contiguous and histologically normal glandular tissue was implied to carry this driver mutation. This suggested that the presence of driver mutations can result in substantial clonal expansion within normal prostatic ductal tissue during adult tissue homeostasis.

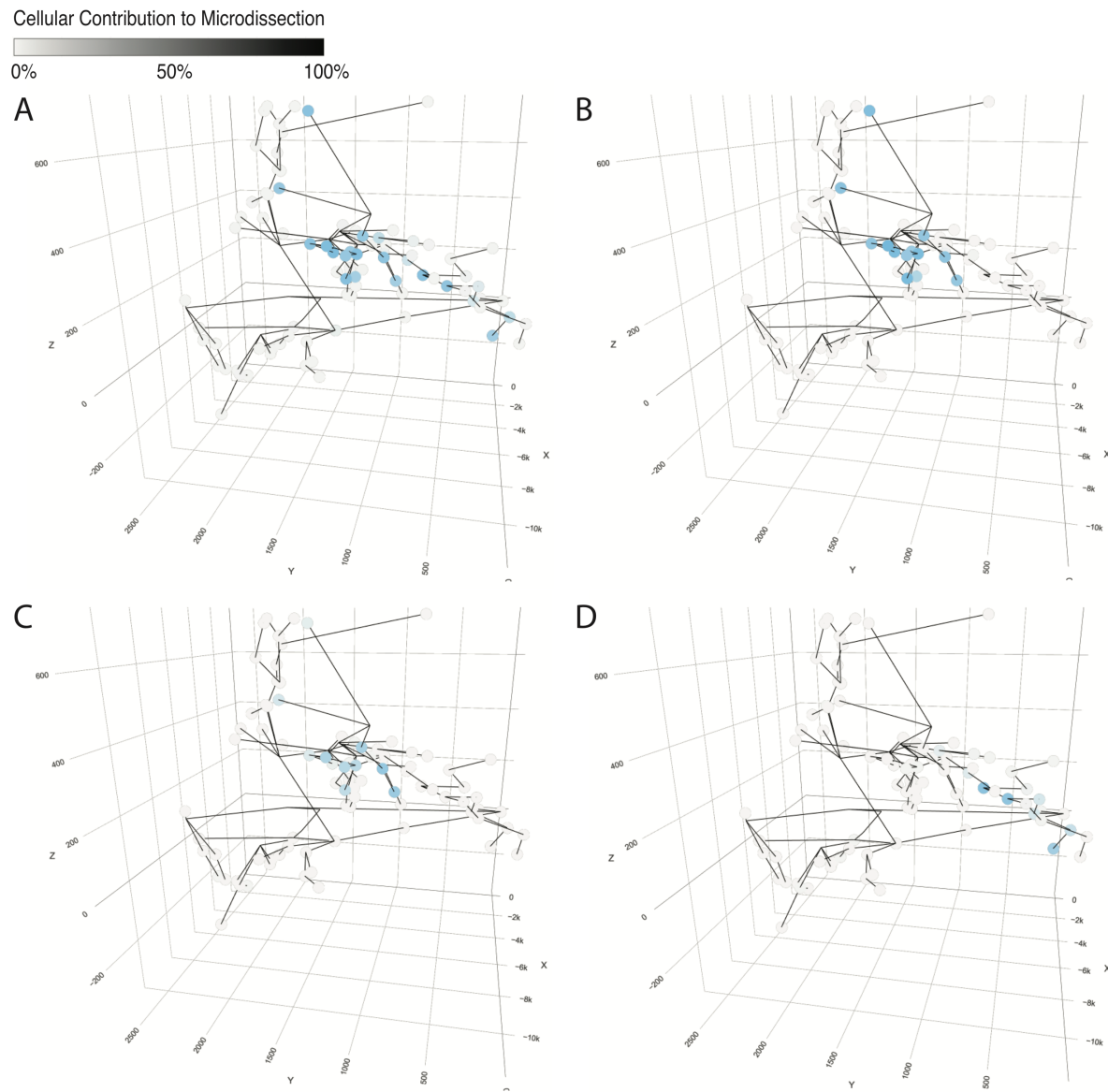


Figure 52: An Adult Clone Carrying a Prostate Cancer Driver Mutation Displays Great Spatial Distribution Within Normal Glandular Epithelium. The lineage tree for the corresponding glandular subunit on the right-hand side is shown in Figure 43. **(A)** The embryonic clone represented by Cluster 98 displays wide spatial distribution in two main peripheral ducts. **(B)** Cluster 118 associated with a *FOXA1* driver mutation represents an adult clone that occupies most of the territory of its corresponding ancestral clone that existed during embryonic development. **(C)** Cluster 36 is subclonal to Cluster 118. **(D)** The pubertal clone represented by Cluster 85 is a parallel branch to Cluster 118 and is shown for comparison of clonal expansion during pubertal morphogenesis.

SBS40 is Only Detected in Recent Ancestral Clones

The mutational landscape in normal prostatic epithelium was shown to be dominated by clock-like mutational signatures. Notably, mutational signatures were slightly more diverse in microdissections from older prostate tissue donors. SBS40 was the only

additional signature identified in prostatic tissue from two donors with higher contribution in the older donor (Figure 41). In addition to comparison of the mutational signatures across donors, the phylogenetic trees of glandular subunits allowed for the investigation of changing mutational signatures with time within the same patient. Mutational signatures were extracted for all clusters that could be placed in the phylogenetic trees and SBS1, SBS5 and SBS40 were identified for both glandular subunits. Notably, SBS40 was only identified in two terminal branches for the glandular subunit on the left-hand side and in eleven terminal branches for the respective subunit on the right-hand side of the prostate. In these thirteen terminal branches, SBS40 contributed to a median of 50% of the associated mutations with SBS1 accounting for a median of 11% and SBS5 for a median of 39%. For the remaining branches, median contributions of SBS1 of 14% and SBS5 of 86% were observed (Figure 53). Hence, SBS40 was only observed in more recent clones that were not associated with embryonic or pubertal development. However, not all terminal branches displayed a contribution of SBS40, indicating a more complex relationship than the switch from development to tissue maintenance.

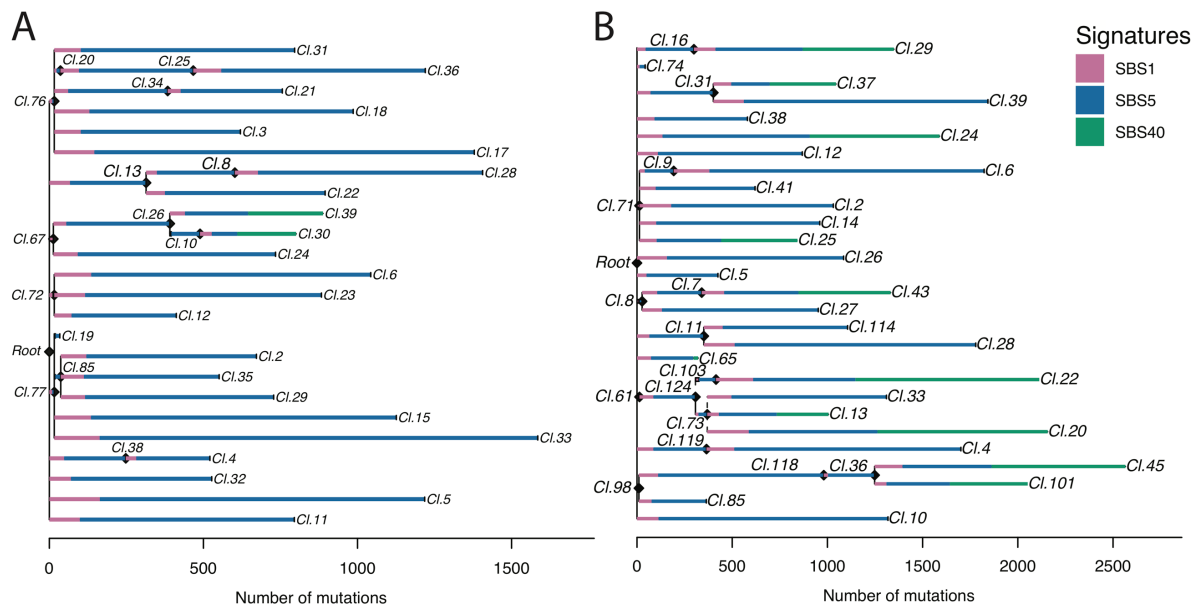


Figure 53: Contribution of SBS40 is Only Detected in More Recent Common Ancestors. Mutational signatures were extracted for all ancestral clones that were placed on the lineage trees for two glandular subunits. The clock-like mutational signatures SBS1, SBS5 and SBS40 were identified in both structures. Notably, SBS40 was only detected in more terminal branches. These terminal branches were associated with more recent common ancestors that have existed during adult tissue homeostasis.

Discussion

The presented work highlights multiple insights into the mutational landscape of normal prostate epithelium with relevance to prostate cancer. Mutations were demonstrated to accumulate with age as well as throughout the ductal network of individual glandular subunits with increasing distance from the peri-urethral ducts (Figure 40 and 45A). These observations align well with the strongly increased prostate cancer risk in older men in general and in glandular acini in particular as these represent the most distal parts of the glandular network ^{307,311}.

The spatial effect of a proximal to distal increase in mutational burden could only be considered for the donor of the whole prostate sample, where morphological reconstruction of individual glandular subunits was possible (Figure 42 - 44). For the remaining donors, small prostate biopsies were obtained without information about their original anatomic location. While primarily small terminal ducts and acini were sampled for WGS for these donors, a precise localisation within individual glandular subunits was impossible. To improve the estimates for age-dependent accumulation of mutation, sampling within morphologically reconstructed glandular tissue would be essential to account for the substantial intra-individual variation observed in the presented data set. Despite these limitations, a significant linear correlation of mutation accumulation with age could be shown (Figure 40). The proposed model estimated an increase of 16 mutations per year with an offset of 150 mutations. Notably, the considered age range was between 22 and 78 years of age. Therefore, all considered donors had completed pubertal development, which is associated with extensive morphogenesis ^{299,300,303-305}. Extending the considered age range to pre-pubertal stages would allow to estimate a potential difference in mutation rates before and after finalisation of prostatic development, which is suggested by the substantial offset. However, access to pre-pubertal human prostate samples is extremely limited. Consequently, the estimated mutation rate of 16 per year should be limited to the phase of adult tissue maintenance.

This mutation rate is lower than in several other adult human tissues such as small intestine, colon, liver, endometrium or oesophagus with estimated rates between 23 – 42 mutations per year ^{68,70,84,92}. However, it is well in line with the mutation rates for satellite cells in skeletal muscle, which were estimated to acquire 13 mutations per

year ²⁷⁴. Remarkably, both satellite cells in skeletal muscle as well as adult prostate epithelium are quiescent cell and tissue types ^{324,345}. Therefore, the lower mutation rates in these tissues potentially mirror a reduced cellular turnover.

The pronounced interdependence between cell division and mutation accumulation in normal glandular epithelium of the prostate was suggested by the significant relationship between telomere length and mutational burden as well as the dominance of clock-like mutational processes (Figure 45B and 41) ⁶⁶. SBS1 and SBS5 contributed to the majority of mutations in all donors and SBS40 in two of the older donors (Figure 41). Remarkably, presence of SBS40 could only be detected in terminal branches on the lineage trees that represent more recent clones for both morphologically reconstructed glandular subunits (Figure 53). Collectively, these observations suggest that additional mutational processes are active in the aging prostate. Given the strong age-dependence of prostate cancer and observation of SBS40 in corresponding genomes, this highlights a potential disease relevance of these mutational processes ³⁶.

In addition to the insights into the mutational landscape of normal prostatic glandular epithelium, the observed mutations and their distribution across individual glandular subunits allowed for exploration of the clonal dynamics within the human prostate. Individual mutations were clustered to identify ancestral clones that could be placed onto phylogenetic trees (Figure 46 - 48). Using information about the temporal sequence of existence of these clones and their corresponding mutational burden, clones were associated with embryonic and pubertal development as well as with adult tissue maintenance. Through combination of this phylogenetic information with the spatial distribution of these clones within morphologically reconstructed glandular subunits, several features of development and tissue maintenance could be inferred (Figure 49 - 52).

Only few embryonic clones could be observed but these provided high cellular contributions to nearly all microdissections sampled throughout the ductal network (Figure 49). This highlights that only small numbers, maybe as few as four to six, embryonic cells give rise to whole glandular subunits during embryonic development. Notably, embryonic clones displayed a contiguous distribution throughout the ductal

network and could be found in proximity to other embryonic clones. However, their contribution to individual microdissections was largely mutually exclusive (Figure 49). These findings are in line with anatomic observations during embryonic development and corroborate the suggested stochasticity during branching morphogenesis^{290,301,302}. Budding tips extend from the urethra-proximal prostate to generate the immature ductal network and the cells present at budding tip formation give rise to the whole glandular subunit. Stochastic branching of a subset of the cells within each budding tip explain the overall wide distribution throughout the glandular subunit but mutually exclusive contribution to smaller terminal side branches. Such a stochastic and long-ranging model of clonal distribution during embryonic development is also in line with the *in-situ* lineage tracking of COX-deficient clones in prostatic glandular epithelium^{295,331,332}.

Following the phase of embryonic development, a second group of ancestral clones was associated with pubertal development in accordance to the model considering age-related accumulation of mutations and through similarity of their mutational burden with the one observed in microdissections from a 22-year-old donor (Figure 40). Notably, the bimodal clustering of coalescences implied to occur during embryonic and pubertal development, respectively, aligns well with the described quiescence between these periods and distinct morphogenic bursts^{299,303,305}. Pubertal clones displayed a contiguous but more spatially confined distribution than embryonic clones. When a pubertal clone could be associated with an embryonic clone, the pubertal clone was always observed within the area covered by the embryonic clone. Some pubertal clones covered several sampled side ducts but their corresponding branching points were always in very close distance to each other (Figure 49 and 50). Commonly, the cellular contribution of pubertal clones implied at least bipotent lineage potential. These observations can be explained through the second burst of branching morphogenesis during adolescence^{299,303,305}. Budding tips are still observed throughout the ductal network of the pubertal prostate and are enriched at branching points³⁰⁰. Directed morphogenesis in proximal to distal direction similar to embryonic development gives rise to the more complex adult glandular network^{299,300,303}. However, the presented data implies that scattered stem cells within budding tips can initiate this branching morphogenesis throughout the ductal network and contribution from a peri-urethral stem cell compartment is not strictly necessary.

Subsequent to adolescence, the prostate enters a stage of adult tissue homeostasis. Notably, the human prostate is a comparatively quiescent organ and the time for epithelial cell turnover is estimated to be around 200 days ³²⁴. Despite the low proliferation rates, several clones that have existed during the period of adult tissue maintenance could be identified. These clones displayed an even more spatially confined distribution than pubertal clones and significant cellular contribution was usually only observed in one or few adjacent microdissections. Remarkably, despite the long cellular turnover times, adult clones featured hundreds of private mutations that were commonly not detectable in adjacent microdissections. Cellular contribution of adult clones often increased in proximal to distal direction but few examples of the most substantial contribution in more proximal microdissections were also observed. The cellular contribution for several of these clones indicated at least a bipotent lineage potential (Figure 51).

The observation that hundreds of mutations are acquired in common ancestors for basal and luminal epithelium across a distance of a few hundred micrometres, suggests a localised population of cells that is long-lived, has high proliferative capacity and multipotent lineage potential. This aligns well with the previous identification of small and isolated cell populations that were scattered throughout the ductal network and display stem cell markers ³²⁴⁻³²⁸. While the previously published model of continuous directed migration of basal stem cells from a peri-urethral niche to distal parts of the glandular network cannot be ruled out with the presented data, any directed migration is suggested to be an extremely slow process ²⁹⁵. Nearly the full mutational burden per microdissection that is dated to adult tissue maintenance is private to a ductal area of few hundred micrometres. Therefore, the presented data primarily supports a model of scattered ASCs during tissue maintenance and the detection of directed long-ranging COX-deficient clones would be attributed to morphogenesis during pre- and post-natal development. To further corroborate this model, a more precise and location-specific determination of the epithelial cell turnover time and the corresponding mutation rate would be needed. Additionally, targeted deep-sequencing of the previously identified mutations and more extensive microdissection sampling throughout the ductal system would allow for a more precise definition of clonal territories to estimate the contribution of residual directed migration during the adult life stage more accurately.

A model of local proliferation of a stem-cell like population during adult tissue maintenance is further supported by the fact that the only adult clones that displayed greater spatial distribution were linked to a prostate cancer driver mutation. Cluster 118 included the R219S mutation in *FOXA1* and displayed greater spatial distribution than a parallel pubertal clone that originated from the same embryonic clone (Figure 52). *FOXA1* mutations are used as molecular classification marker in prostate cancer and were demonstrated to substantially increase cellular proliferation rates³⁴¹⁻³⁴³. Cluster 118 was observed throughout 13 contiguous microdissections covering a main peripheral duct and several corresponding side branches (Figure 52). Notably, substantial subclonal evolution within this local area was suggested by three further clones that derived from cellular progeny of Cluster 118 (Figure 48 and 52). This extensive subclonal evolution during the adult life stage within a localised area of a peripheral duct is another strong indication of a long-lived but localised cell population that is responsible for tissue maintenance.

Furthermore, the spatial expansion of a clone carrying a proliferation-enhancing prostate cancer driver mutation within normal glandular epithelium highlights the selective advantage of these mutations within normal tissues^{137,343}. While high prevalence of driver mutation-associated clones was demonstrated within normal skin, oesophagus and endometrium, clonal expansion of the *FOXA1*-associated clone was the only one observed in the presented data and no global signals of positive selection could be detected^{68-70,92,93}. However, the mutational burden in prostate cancer is comparatively low, the driver landscape heterogeneous and only the most prevalent rearrangements were considered in this study^{307,317,319}. Therefore, additional disease-relevant mutations could have been missed. Still, the remaining clonal expansions were potentially linked to developmental morphogenesis, which supports a primarily neutral genomic background in histologically normal glandular epithelium of the human prostate.

The proposed model can be summarised as directed morphogenesis during development followed by an adult phase dominated by proliferation of scattered progenitor cells with rare expansions of driver mutation-associated clones that can display extensive subclonal evolution. This model aligns the previously reported observations of distinct stem cell populations throughout the ductal networks with the

long-ranging COX-deficient clones reported from *in-situ* lineage tracking ^{295,324-328,331,332}. Moreover, it provides an explanation for the observation of multiple prostate cancer foci, which are implied to share some of their mutational history but are suggested to have eventually derived from distinct clones ³¹³⁻³¹⁶.

In summary, the presented work highlights mutational processes in normal glandular epithelium that are relevant to our understanding of prostate cancer development. Moreover, this is the first description of clonal dynamics within prostatic glandular subunits based on a whole-genome scale. Differences between developmental and tissue maintenance stages are implied and a model that aligns previous insights into human prostate stem cell biology and disease-relevant features is proposed.

Conclusion

Somatic mutation is a natural process that starts at fertilization and continues to shape tissue and cell heterogeneity throughout life. In the presented thesis, various techniques are advanced or applied to study the relevance of somatic mutation in the context of health and disease. Furthermore, the continuity of somatic mutation is utilised to retrospectively analyse the clonal dynamics within a human tissue, avoiding the need for artificial markers or model organisms.

The mutational landscapes of human tissues are diverse and can instruct us about their cellular biology. Here, the mutational spectrum of somatic base substitutions was reported for cortical neurons that exist in a post-mitotic state for several decades and a direct link to active transcription was established. In contrast, the mutational burden acquired through substantial cellular proliferation during two distinct morphogenic bursts identified potentially neutral and long-ranging clonal expansion in the human prostate during embryonic and pubertal development.

While somatic mutation is common in normal human tissue, distinct alterations or combinations thereof can lead to severe tissue malfunction and disease. Recent surveys have started to explore the mutational processes in the corresponding normal tissues and most of the work presented in this thesis is placed within this context. While especially meta-analyses of the rich data available for the diseased state will continue to further our understanding of aberrant processes as demonstrated in one of the presented chapters, deeper insights into the mutational landscape and somatic evolution within normal tissues are crucial to unravel the origin of disease.

References

- 1 Weinberg, R. A. *The biology of cancer*. (Garland Science, Taylor & Francis Group, 2014).
- 2 Dobzhansky, T. in *Studies in the Philosophy of Biology* (eds Francisco Ayala & Theodosius Dobzhansky) (Palgrave, 1974).
- 3 Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860-921, doi:10.1038/35057062 (2001).
- 4 Venter, J. C. *et al.* The sequence of the human genome. *Science* **291**, 1304-1351, doi:10.1126/science.1058040 (2001).
- 5 Levy, S. *et al.* The diploid genome sequence of an individual human. *PLoS Biol* **5**, e254, doi:10.1371/journal.pbio.0050254 (2007).
- 6 Karki, R., Pandya, D., Elston, R. C. & Ferlini, C. Defining "mutation" and "polymorphism" in the era of personal genomics. *BMC Med Genomics* **8**, 37, doi:10.1186/s12920-015-0115-z (2015).
- 7 Altshuler, D. *et al.* An SNP map of the human genome generated by reduced representation shotgun sequencing. *Nature* **407**, 513-516, doi:10.1038/35035083 (2000).
- 8 Sachidanandam, R. *et al.* A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409**, 928-933, doi:10.1038/35057149 (2001).
- 9 Strachan, T. & Read, A. *Human Molecular Genetics*. 5th Edition edn, (Garland Science, 2018).
- 10 Lindahl, T. Instability and decay of the primary structure of DNA. *Nature* **362**, 709-715, doi:10.1038/362709a0 (1993).
- 11 Lindahl, T. & Nyberg, B. Rate of depurination of native deoxyribonucleic acid. *Biochemistry* **11**, 3610-3618, doi:10.1021/bi00769a018 (1972).
- 12 Cadet, J. & Wagner, J. R. DNA base damage by reactive oxygen species, oxidizing agents, and UV radiation. *Cold Spring Harb Perspect Biol* **5**, doi:10.1101/cshperspect.a012559 (2013).
- 13 Branzei, D. & Foiani, M. The DNA damage response during DNA replication. *Curr Opin Cell Biol* **17**, 568-575, doi:10.1016/j.ceb.2005.09.003 (2005).
- 14 D'Alessandro, G. & d'Adda di Fagagna, F. Transcription and DNA Damage: Holding Hands or Crossing Swords? *J Mol Biol* **429**, 3215-3229, doi:10.1016/j.jmb.2016.11.002 (2017).
- 15 Loeb, L. A. & Harris, C. C. Advances in chemical carcinogenesis: a historical review and prospective. *Cancer Res* **68**, 6863-6872, doi:10.1158/0008-5472.CAN-08-2852 (2008).
- 16 Rastogi, R. P., Richa, Kumar, A., Tyagi, M. B. & Sinha, R. P. Molecular mechanisms of ultraviolet radiation-induced DNA damage and repair. *J Nucleic Acids* **2010**, 592980, doi:10.4061/2010/592980 (2010).
- 17 Stephens, P. J. *et al.* Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell* **144**, 27-40, doi:10.1016/j.cell.2010.11.055 (2011).
- 18 Nik-Zainal, S. *et al.* Mutational processes molding the genomes of 21 breast cancers. *Cell* **149**, 979-993, doi:10.1016/j.cell.2012.04.024 (2012).
- 19 Lindahl, T. & Barnes, D. E. Repair of endogenous DNA damage. *Cold Spring Harb Symp Quant Biol* **65**, 127-133 (2000).

- 20 Tomasetti, C. & Vogelstein, B. Cancer etiology. Variation in cancer risk among tissues can be explained by the number of stem cell divisions. *Science* **347**, 78-81, doi:10.1126/science.1260825 (2015).
- 21 Wu, S., Powers, S., Zhu, W. & Hannun, Y. A. Substantial contribution of extrinsic risk factors to cancer development. *Nature* **529**, 43-47, doi:10.1038/nature16166 (2016).
- 22 Hoeijmakers, J. H. DNA damage, aging, and cancer. *N Engl J Med* **361**, 1475-1485, doi:10.1056/NEJMra0804615 (2009).
- 23 Tubbs, A. & Nussenzweig, A. Endogenous DNA Damage as a Source of Genomic Instability in Cancer. *Cell* **168**, 644-656, doi:10.1016/j.cell.2017.01.002 (2017).
- 24 Ciccica, A. & Elledge, S. J. The DNA damage response: making it safe to play with knives. *Mol Cell* **40**, 179-204, doi:10.1016/j.molcel.2010.09.019 (2010).
- 25 Chatterjee, N. & Walker, G. C. Mechanisms of DNA damage, repair, and mutagenesis. *Environ Mol Mutagen* **58**, 235-263, doi:10.1002/em.22087 (2017).
- 26 Scully, R., Panday, A., Elango, R. & Willis, N. A. DNA double-strand break repair-pathway choice in somatic mammalian cells. *Nat Rev Mol Cell Biol*, doi:10.1038/s41580-019-0152-0 (2019).
- 27 Roos, W. P. & Kaina, B. DNA damage-induced cell death by apoptosis. *Trends Mol Med* **12**, 440-450, doi:10.1016/j.molmed.2006.07.007 (2006).
- 28 Bianconi, E. *et al.* An estimation of the number of cells in the human body. *Ann Hum Biol* **40**, 463-471, doi:10.3109/03014460.2013.807878 (2013).
- 29 Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500**, 415-421, doi:10.1038/nature12477 (2013).
- 30 Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Campbell, P. J. & Stratton, M. R. Deciphering signatures of mutational processes operative in human cancer. *Cell Rep* **3**, 246-259, doi:10.1016/j.celrep.2012.12.008 (2013).
- 31 Fischer, A., Illingworth, C. J., Campbell, P. J. & Mustonen, V. EMu: probabilistic inference of mutational processes and their localization in the cancer genome. *Genome Biol* **14**, R39, doi:10.1186/gb-2013-14-4-r39 (2013).
- 32 Rosales, R. A., Drummond, R. D., Valieris, R., Dias-Neto, E. & da Silva, I. T. signeR: an empirical Bayesian approach to mutational signature discovery. *Bioinformatics* **33**, 8-16, doi:10.1093/bioinformatics/btw572 (2017).
- 33 Kim, J. *et al.* Somatic ERCC2 mutations are associated with a distinct genomic signature in urothelial tumors. *Nat Genet* **48**, 600-606, doi:10.1038/ng.3557 (2016).
- 34 Nik-Zainal, S. *et al.* Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature* **534**, 47-54, doi:10.1038/nature17676 (2016).
- 35 Li, Y. *et al.* Patterns of structural variation in human cancer. *bioRxiv*, 181339, doi:10.1101/181339 (2017).
- 36 Alexandrov, L. B. *et al.* The Repertoire of Mutational Signatures in Human Cancer. *bioRxiv*, 322859, doi:10.1101/322859 (2018).
- 37 Funnell, T. *et al.* Integrated structural variation and point mutation signatures in cancer genomes using correlated topic models. *PLoS Comput Biol* **15**, e1006799, doi:10.1371/journal.pcbi.1006799 (2019).
- 38 Sturtevant, A. H. Essays on evolution I On the effects of selection on mutation rate. *Q Rev Biol* **12**, 464-467, doi:Doi 10.1086/394543 (1937).

- 39 Kimura, M. Optimum mutation rate and degree of dominance as determined by the principle of minimum genetic load. *Journal of Genetics* **57**, 21-34, doi:10.1007/bf02985336 (1960).
- 40 Kimura, M. The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. *Genetics* **61**, 893-903 (1969).
- 41 Baer, C. F., Miyamoto, M. M. & Denver, D. R. Mutation rate variation in multicellular eukaryotes: causes and consequences. *Nat Rev Genet* **8**, 619-631, doi:10.1038/nrg2158 (2007).
- 42 Conrad, D. F. *et al.* Variation in genome-wide mutation rates within and between human families. *Nat Genet* **43**, 712-714, doi:10.1038/ng.862 (2011).
- 43 Kong, A. *et al.* Rate of de novo mutations and the importance of father's age to disease risk. *Nature* **488**, 471-475, doi:10.1038/nature11396 (2012).
- 44 Kirkwood, T. B. Evolution of ageing. *Nature* **270**, 301-304, doi:10.1038/270301a0 (1977).
- 45 Carlson, E. A. & Southin, J. L. Chemically Induced Somatic and Gonadal Mosaicism in Drosophila. I. Sex-Linked Lethals. *Genetics* **48**, 663-675 (1963).
- 46 Lupski, J. R. Genetics. Genome mosaicism—one human, multiple genomes. *Science* **341**, 358-359, doi:10.1126/science.1239503 (2013).
- 47 Lynch, M. Evolution of the mutation rate. *Trends Genet* **26**, 345-352, doi:10.1016/j.tig.2010.05.003 (2010).
- 48 Gossen, J. A. *et al.* Efficient rescue of integrated shuttle vectors from transgenic mice: a model for studying mutations in vivo. *Proc Natl Acad Sci U S A* **86**, 7971-7975, doi:10.1073/pnas.86.20.7971 (1989).
- 49 Boerriqter, M. E., Dolle, M. E., Martus, H. J., Gossen, J. A. & Vijg, J. Plasmid-based transgenic mouse model for studying in vivo mutations. *Nature* **377**, 657-659, doi:10.1038/377657a0 (1995).
- 50 Shendure, J. & Ji, H. Next-generation DNA sequencing. *Nat Biotechnol* **26**, 1135-1145, doi:10.1038/nbt1486 (2008).
- 51 Dou, Y., Gold, H. D., Luquette, L. J. & Park, P. J. Detecting Somatic Mutations in Normal Cells. *Trends Genet* **34**, 545-557, doi:10.1016/j.tig.2018.04.003 (2018).
- 52 Griffiths, A. J. F., Miller, J. H. & Suzuki, D. in *An Introduction to Genetic Analysis* (W. H. Freeman, 2000).
- 53 Dal, G. M. *et al.* Early postzygotic mutations contribute to de novo variation in a healthy monozygotic twin pair. *J Med Genet* **51**, 455-459, doi:10.1136/jmedgenet-2013-102197 (2014).
- 54 Li, R. *et al.* Somatic point mutations occurring early in development: a monozygotic twin study. *J Med Genet* **51**, 28-34, doi:10.1136/jmedgenet-2013-101712 (2014).
- 55 Morimoto, Y. *et al.* Deep sequencing reveals variations in somatic cell mosaic mutations between monozygotic twins with discordant psychiatric disease. *Hum Genome Var* **4**, 17032, doi:10.1038/hgv.2017.32 (2017).
- 56 Vadlamudi, L. *et al.* Timing of de novo mutagenesis—a twin study of sodium-channel mutations. *N Engl J Med* **363**, 1335-1340, doi:10.1056/NEJMoa0910752 (2010).
- 57 Vogt, J. *et al.* Monozygotic twins discordant for neurofibromatosis type 1 due to a postzygotic NF1 gene mutation. *Hum Mutat* **32**, E2134-2147, doi:10.1002/humu.21476 (2011).

- 58 Nishioka, M. *et al.* Identification of somatic mutations in monozygotic twins discordant for psychiatric disorders. *NPJ Schizophr* **4**, 7, doi:10.1038/s41537-018-0049-5 (2018).
- 59 Bruder, C. E. *et al.* Phenotypically concordant and discordant monozygotic twins display different DNA copy-number-variation profiles. *Am J Hum Genet* **82**, 763-771, doi:10.1016/j.ajhg.2007.12.011 (2008).
- 60 Ju, Y. S. *et al.* Somatic mutations reveal asymmetric cellular dynamics in the early human embryo. *Nature* **543**, 714-718, doi:10.1038/nature21703 (2017).
- 61 Kuijk, E. *et al.* Early divergence of mutational processes in human fetal tissues. *Sci Adv* **5**, eaaw1271, doi:10.1126/sciadv.aaw1271 (2019).
- 62 Genovese, G. *et al.* Clonal hematopoiesis and blood-cancer risk inferred from blood DNA sequence. *N Engl J Med* **371**, 2477-2487, doi:10.1056/NEJMoa1409405 (2014).
- 63 Jaiswal, S. *et al.* Age-related clonal hematopoiesis associated with adverse outcomes. *N Engl J Med* **371**, 2488-2498, doi:10.1056/NEJMoa1408617 (2014).
- 64 Holstege, H. *et al.* Somatic mutations found in the healthy blood compartment of a 115-yr-old woman demonstrate oligoclonal hematopoiesis. *Genome Res* **24**, 733-742, doi:10.1101/gr.162131.113 (2014).
- 65 Milholland, B., Auton, A., Suh, Y. & Vijg, J. Age-related somatic mutations in the cancer genome. *Oncotarget* **6**, 24627-24635, doi:10.18632/oncotarget.5685 (2015).
- 66 Alexandrov, L. B. *et al.* Clock-like mutational processes in human somatic cells. *Nat Genet* **47**, 1402-1407, doi:10.1038/ng.3441 (2015).
- 67 Saini, N. & Gordenin, D. A. Somatic mutation load and spectra: A record of DNA damage and repair in healthy human cells. *Environ Mol Mutagen* **59**, 672-686, doi:10.1002/em.22215 (2018).
- 68 Martincorena, I. *et al.* Somatic mutant clones colonize the human esophagus with age. *Science* **362**, 911-917, doi:10.1126/science.aau3879 (2018).
- 69 Martincorena, I. *et al.* Tumor evolution. High burden and pervasive positive selection of somatic mutations in normal human skin. *Science* **348**, 880-886, doi:10.1126/science.aaa6806 (2015).
- 70 Yokoyama, A. *et al.* Age-related remodelling of oesophageal epithelia by mutated cancer drivers. *Nature* **565**, 312-317, doi:10.1038/s41586-018-0811-x (2019).
- 71 Kinde, I., Wu, J., Papadopoulos, N., Kinzler, K. W. & Vogelstein, B. Detection and quantification of rare mutations with massively parallel sequencing. *Proc Natl Acad Sci U S A* **108**, 9530-9535, doi:10.1073/pnas.1105422108 (2011).
- 72 Kennedy, S. R. *et al.* Detecting ultralow-frequency mutations by Duplex Sequencing. *Nat Protoc* **9**, 2586-2606, doi:10.1038/nprot.2014.170 (2014).
- 73 Schmitt, M. W. *et al.* Detection of ultra-rare mutations by next-generation sequencing. *Proc Natl Acad Sci U S A* **109**, 14508-14513, doi:10.1073/pnas.1208715109 (2012).
- 74 Gregory, M. T. *et al.* Targeted single molecule mutation detection with massively parallel sequencing. *Nucleic Acids Res* **44**, e22, doi:10.1093/nar/gkv915 (2016).
- 75 Hoang, M. L. *et al.* Genome-wide quantification of rare somatic mutations in normal human tissues using massively parallel sequencing. *Proc Natl Acad Sci U S A* **113**, 9846-9851, doi:10.1073/pnas.1607794113 (2016).

- 76 Abyzov, A. *et al.* Somatic copy number mosaicism in human skin revealed by induced pluripotent stem cells. *Nature* **492**, 438-442, doi:10.1038/nature11629 (2012).
- 77 Abyzov, A. *et al.* One thousand somatic SNVs per skin fibroblast cell set baseline of mosaic mutational load with patterns that suggest proliferative origin. *Genome Res* **27**, 512-523, doi:10.1101/gr.215517.116 (2017).
- 78 Friedmann-Morvinski, D. *et al.* Dedifferentiation of neurons and astrocytes by oncogenes can induce gliomas in mice. *Science* **338**, 1080-1084, doi:10.1126/science.1226929 (2012).
- 79 Rouhani, F. J. *et al.* Mutational History of a Human Cell Lineage from Somatic to Induced Pluripotent Stem Cells. *PLoS Genet* **12**, e1005932, doi:10.1371/journal.pgen.1005932 (2016).
- 80 Yamazaki, Y. *et al.* Assessment of the developmental totipotency of neural cells in the cerebral cortex of mouse embryo by nuclear transfer. *Proc Natl Acad Sci U S A* **98**, 14022-14026, doi:10.1073/pnas.231489398 (2001).
- 81 Hazen, J. L. *et al.* The Complete Genome Sequences, Unique Mutational Spectra, and Developmental Potency of Adult Neurons Revealed by Cloning. *Neuron* **89**, 1223-1236, doi:10.1016/j.neuron.2016.02.004 (2016).
- 82 Behjati, S. *et al.* Genome sequencing of normal cells reveals developmental lineages and mutational processes. *Nature* **513**, 422-425, doi:10.1038/nature13448 (2014).
- 83 Saini, N. *et al.* The Impact of Environmental and Endogenous Damage on Somatic Mutation Load in Human Skin Fibroblasts. *PLoS Genet* **12**, e1006385, doi:10.1371/journal.pgen.1006385 (2016).
- 84 Blokzijl, F. *et al.* Tissue-specific mutation accumulation in human adult stem cells during life. *Nature* **538**, 260-264, doi:10.1038/nature19768 (2016).
- 85 Welch, J. S. *et al.* The origin and evolution of mutations in acute myeloid leukemia. *Cell* **150**, 264-278, doi:10.1016/j.cell.2012.06.023 (2012).
- 86 Lee-Six, H. *et al.* Population dynamics of normal human blood inferred from somatic mutations. *Nature* **561**, 473-478, doi:10.1038/s41586-018-0497-0 (2018).
- 87 Bae, T. *et al.* Different mutational rates and mechanisms in human cells at pregastrulation and neurogenesis. *Science* **359**, 550-555, doi:10.1126/science.aan8690 (2018).
- 88 Datta, S. *et al.* Laser capture microdissection: Big data from small samples. *Histol Histopathol* **30**, 1255-1269, doi:10.14670/HH-11-622 (2015).
- 89 Griffiths, D. F., Davies, S. J., Williams, D., Williams, G. T. & Williams, E. D. Demonstration of somatic mutation and colonic crypt clonality by X-linked enzyme histochemistry. *Nature* **333**, 461-463, doi:10.1038/333461a0 (1988).
- 90 Nicholson, A. M. *et al.* Fixation and Spread of Somatic Mutations in Adult Human Colonic Epithelium. *Cell Stem Cell* **22**, 909-918 e908, doi:10.1016/j.stem.2018.04.020 (2018).
- 91 Lee-Six, H. *et al.* The landscape of somatic mutation in normal colorectal epithelial cells. *bioRxiv*, 416800, doi:10.1101/416800 (2018).
- 92 Moore, L. *et al.* The mutational landscape of normal human endometrial epithelium. *bioRxiv*, 505685, doi:10.1101/505685 (2018).
- 93 Suda, K. *et al.* Clonal Expansion and Diversification of Cancer-Associated Mutations in Endometriosis and Normal Endometrium. *Cell Rep* **24**, 1777-1789, doi:10.1016/j.celrep.2018.07.037 (2018).

- 94 Knouse, K. A., Wu, J. & Amon, A. Assessment of megabase-scale somatic copy number variation using single-cell sequencing. *Genome Res* **26**, 376-384, doi:10.1101/gr.198937.115 (2016).
- 95 Knouse, K. A., Wu, J., Whittaker, C. A. & Amon, A. Single cell sequencing reveals low levels of aneuploidy across mammalian tissues. *Proc Natl Acad Sci U S A* **111**, 13409-13414, doi:10.1073/pnas.1415287111 (2014).
- 96 McConnell, M. J. *et al.* Mosaic copy number variation in human neurons. *Science* **342**, 632-637, doi:10.1126/science.1243472 (2013).
- 97 Cai, X. *et al.* Single-cell, genome-wide sequencing identifies clonal somatic copy-number variation in the human brain. *Cell Rep* **8**, 1280-1289, doi:10.1016/j.celrep.2014.07.043 (2014).
- 98 Evrony, G. D. *et al.* Single-neuron sequencing analysis of L1 retrotransposition and somatic mutation in the human brain. *Cell* **151**, 483-496, doi:10.1016/j.cell.2012.09.035 (2012).
- 99 Evrony, G. D. *et al.* Cell lineage analysis in human brain using endogenous retroelements. *Neuron* **85**, 49-59, doi:10.1016/j.neuron.2014.12.028 (2015).
- 100 Evrony, G. D., Lee, E., Park, P. J. & Walsh, C. A. Resolving rates of mutation in the brain using single-neuron genomics. *Elife* **5**, doi:10.7554/eLife.12966 (2016).
- 101 Upton, K. R. *et al.* Ubiquitous L1 mosaicism in hippocampal neurons. *Cell* **161**, 228-239, doi:10.1016/j.cell.2015.03.026 (2015).
- 102 Lodato, M. A. *et al.* Aging and neurodegeneration are associated with increased mutations in single human neurons. *Science* **359**, 555-559, doi:10.1126/science.aao4426 (2018).
- 103 Lodato, M. A. *et al.* Somatic mutation in single human neurons tracks developmental and transcriptional history. *Science* **350**, 94-98, doi:10.1126/science.aab1785 (2015).
- 104 Dong, X. *et al.* Accurate identification of single nucleotide variants in whole genome amplified single cells. *Cancer Research* **77**, doi:10.1158/1538-7445.Am2017-5400 (2017).
- 105 Huang, L., Ma, F., Chapman, A., Lu, S. & Xie, X. S. Single-Cell Whole-Genome Amplification and Sequencing: Methodology and Applications. *Annu Rev Genomics Hum Genet* **16**, 79-102, doi:10.1146/annurev-genom-090413-025352 (2015).
- 106 Hou, Y. *et al.* Comparison of variations detection between whole-genome amplification methods used in single-cell resequencing. *Gigascience* **4**, 37, doi:10.1186/s13742-015-0068-3 (2015).
- 107 Zhang, L. & Vijg, J. Somatic Mutagenesis in Mammals and Its Implications for Human Disease and Aging. *Annu Rev Genet* **52**, 397-419, doi:10.1146/annurev-genet-120417-031501 (2018).
- 108 Stratton, M. R., Campbell, P. J. & Futreal, P. A. The cancer genome. *Nature* **458**, 719-724, doi:10.1038/nature07943 (2009).
- 109 Hanseemann, D. Ueber asymmetrische Zelltheilung in Epithelkrebsen und deren biologische Bedeutung. *Archiv f. pathol. Anat.* **119**, doi:<https://doi.org/10.1007/BF01882039> (1890).
- 110 Boveri, T. *Zur Frage der Entstehung Maligner Tumoren*. (Gustav Fischer, 1914).
- 111 Rowley, J. D. Letter: A new consistent chromosomal abnormality in chronic myelogenous leukaemia identified by quinacrine fluorescence and Giemsa staining. *Nature* **243**, 290-293, doi:10.1038/243290a0 (1973).

- 112 Nowell, P. C. & Hungerford, D. A. A minute chromosome in human chronic granulocytic leukemia. *Science* **142** (1960).
- 113 Kang, Z. J. *et al.* The Philadelphia chromosome in leukemogenesis. *Chin J Cancer* **35**, 48, doi:10.1186/s40880-016-0108-0 (2016).
- 114 Pon, J. R. & Marra, M. A. Driver and passenger mutations in cancer. *Annu Rev Pathol* **10**, 25-50, doi:10.1146/annurev-pathol-012414-040312 (2015).
- 115 Lawrence, M. S. *et al.* Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature* **505**, 495-501, doi:10.1038/nature12912 (2014).
- 116 Davoli, T. *et al.* Cumulative haploinsufficiency and triplosensitivity drive aneuploidy patterns and shape the cancer genome. *Cell* **155**, 948-962, doi:10.1016/j.cell.2013.10.011 (2013).
- 117 Martincorena, I. *et al.* Universal Patterns of Selection in Cancer and Somatic Tissues. *Cell* **171**, 1029-1041 e1021, doi:10.1016/j.cell.2017.09.042 (2017).
- 118 Knudson, A. G., Jr. Mutation and cancer: statistical study of retinoblastoma. *Proc Natl Acad Sci U S A* **68**, 820-823, doi:10.1073/pnas.68.4.820 (1971).
- 119 Friend, S. H. *et al.* A human DNA segment with properties of the gene that predisposes to retinoblastoma and osteosarcoma. *Nature* **323**, 643-646, doi:10.1038/323643a0 (1986).
- 120 Martin, G. S. Rous sarcoma virus: a function required for the maintenance of the transformed state. *Nature* **227**, 1021-1023, doi:10.1038/2271021a0 (1970).
- 121 Stehelin, D., Varmus, H. E., Bishop, J. M. & Vogt, P. K. DNA related to the transforming gene(s) of avian sarcoma viruses is present in normal avian DNA. *Nature* **260**, 170-173, doi:10.1038/260170a0 (1976).
- 122 Krontiris, T. G. & Cooper, G. M. Transforming activity of human tumor DNAs. *Proc Natl Acad Sci U S A* **78**, 1181-1184, doi:10.1073/pnas.78.2.1181 (1981).
- 123 Shih, C., Padhy, L. C., Murray, M. & Weinberg, R. A. Transforming genes of carcinomas and neuroblastomas introduced into mouse fibroblasts. *Nature* **290**, 261-264, doi:10.1038/290261a0 (1981).
- 124 International Cancer Genome, C. *et al.* International network of cancer genome projects. *Nature* **464**, 993-998, doi:10.1038/nature08987 (2010).
- 125 Cancer Genome Atlas Research, N. *et al.* The Cancer Genome Atlas Pan-Cancer analysis project. *Nat Genet* **45**, 1113-1120, doi:10.1038/ng.2764 (2013).
- 126 Campbell, P. J., Getz, G., Stuart, J. M., Korb, J. O. & Stein, L. D. Pan-cancer analysis of whole genomes. *bioRxiv*, 162784, doi:10.1101/162784 (2017).
- 127 Tate, J. G. *et al.* COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Res* **47**, D941-D947, doi:10.1093/nar/gky1015 (2019).
- 128 Campbell, I. M., Shaw, C. A., Stankiewicz, P. & Lupski, J. R. Somatic mosaicism: implications for disease and transmission genetics. *Trends Genet* **31**, 382-392, doi:10.1016/j.tig.2015.03.013 (2015).
- 129 Nocturne, G. & Mariette, X. Advances in understanding the pathogenesis of primary Sjogren's syndrome. *Nat Rev Rheumatol* **9**, 544-556, doi:10.1038/nrrheum.2013.110 (2013).
- 130 Lumbroso, S., Paris, F., Sultan, C. & European Collaborative, S. Activating Gsalpha mutations: analysis of 113 patients with signs of McCune-Albright syndrome--a European Collaborative Study. *J Clin Endocrinol Metab* **89**, 2107-2113, doi:10.1210/jc.2003-031225 (2004).

- 131 Priest, J. R. *et al.* Early somatic mosaicism is a rare cause of long-QT syndrome. *Proc Natl Acad Sci U S A* **113**, 11555-11560, doi:10.1073/pnas.1607187113 (2016).
- 132 Lindhurst, M. J. *et al.* A mosaic activating mutation in AKT1 associated with the Proteus syndrome. *N Engl J Med* **365**, 611-619, doi:10.1056/NEJMoa1104017 (2011).
- 133 Szilard, L. On the Nature of the Aging Process. *Proc Natl Acad Sci U S A* **45**, 30-45, doi:10.1073/pnas.45.1.30 (1959).
- 134 Morley, A. A. The somatic mutation theory of ageing. *Mutat Res* **338**, 19-23, doi:10.1016/0921-8734(95)00007-s (1995).
- 135 de Grey, A. D. Protagonistic pleiotropy: Why cancer may be the only pathogenic effect of accumulating nuclear mutations and epimutations in aging. *Mech Ageing Dev* **128**, 456-459, doi:10.1016/j.mad.2007.05.005 (2007).
- 136 Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: the next generation. *Cell* **144**, 646-674, doi:10.1016/j.cell.2011.02.013 (2011).
- 137 Martincorena, I. Somatic mutation and clonal expansions in human tissues. *Genome Med* **11**, 35, doi:10.1186/s13073-019-0648-4 (2019).
- 138 Mitelman, F., Johansson, B. & Mertens, F. The impact of translocations and gene fusions on cancer causation. *Nat Rev Cancer* **7**, 233-245, doi:10.1038/nrc2091 (2007).
- 139 Alberts, B., Wilson, J. H. & Hunt, T. *Molecular biology of the cell*. 5th edn, (Garland Science, 2008).
- 140 Zheng, J. Oncogenic chromosomal translocations and human cancer (review). *Oncol Rep* **30**, 2011-2019, doi:10.3892/or.2013.2677 (2013).
- 141 Cancer Genome Atlas, N. Comprehensive molecular characterization of human colon and rectal cancer. *Nature* **487**, 330-337, doi:10.1038/nature11252 (2012).
- 142 Baca, S. C. *et al.* Punctuated evolution of prostate cancer genomes. *Cell* **153**, 666-677, doi:10.1016/j.cell.2013.03.021 (2013).
- 143 Gutierrez, S. *et al.* in *Myeloid Leukemia - Basic Mechanisms of Leukemogenesis* (Intech Open, 2011).
- 144 Grewal, S. I. & Jia, S. Heterochromatin revisited. *Nat Rev Genet* **8**, 35-46, doi:10.1038/nrg2008 (2007).
- 145 Rippe, K. *Genome Organization and Function in the Human Nucleus*. (Wiley-VCH, 2012).
- 146 Erdel, F., Muller-Ott, K. & Rippe, K. Establishing epigenetic domains via chromatin-bound histone modifiers. *Ann N Y Acad Sci* **1305**, 29-43, doi:10.1111/nyas.12262 (2013).
- 147 Disteche, C. M. Dosage compensation of the sex chromosomes and autosomes. *Semin Cell Dev Biol* **56**, 9-18, doi:10.1016/j.semcdb.2016.04.013 (2016).
- 148 Lyon, M. F. Gene action in the X-chromosome of the mouse (*Mus musculus* L.). *Nature* **190**, 372-373, doi:10.1038/190372a0 (1961).
- 149 Migeon, B. R. Choosing the Active X: The Human Version of X Inactivation. *Trends Genet* **33**, 899-909, doi:10.1016/j.tig.2017.09.005 (2017).
- 150 Brown, C. J. *et al.* A gene from the region of the human X inactivation centre is expressed exclusively from the inactive X chromosome. *Nature* **349**, 38-44, doi:10.1038/349038a0 (1991).

- 151 Payer, B. & Lee, J. T. X chromosome dosage compensation: how mammals keep the balance. *Annu Rev Genet* **42**, 733-772, doi:10.1146/annurev.genet.42.110807.091711 (2008).
- 152 Disteche, C. M. & Berletch, J. B. X-chromosome inactivation and escape. *J Genet* **94**, 591-599 (2015).
- 153 Carrel, L. & Willard, H. F. X-inactivation profile reveals extensive variability in X-linked gene expression in females. *Nature* **434**, 400-404, doi:10.1038/nature03479 (2005).
- 154 Csankovszki, G., Panning, B., Bates, B., Pehrson, J. R. & Jaenisch, R. Conditional deletion of Xist disrupts histone macroH2A localization but not maintenance of X inactivation. *Nat Genet* **22**, 323-324, doi:10.1038/11887 (1999).
- 155 Bala Tannan, N. *et al.* DNA methylation profiling in X;autosome translocations supports a role for L1 repeats in the spread of X chromosome inactivation. *Hum Mol Genet* **23**, 1224-1236, doi:10.1093/hmg/ddt553 (2014).
- 156 Giorda, R. *et al.* Molecular and cytogenetic analysis of the spreading of X inactivation in a girl with microcephaly, mild dysmorphic features and t(X;5)(q22.1;q31.1). *Eur J Hum Genet* **16**, 897-905, doi:10.1038/ejhg.2008.28 (2008).
- 157 Jones, C. *et al.* Bilateral retinoblastoma in a male patient with an X; 13 translocation: evidence for silencing of the RB1 gene by the spreading of X inactivation. *Am J Hum Genet* **60**, 1558-1562, doi:10.1016/S0002-9297(07)64254-2 (1997).
- 158 Sakazume, S. *et al.* Spread of X-chromosome inactivation into chromosome 15 is associated with Prader-Willi syndrome phenotype in a boy with a t(X;15)(p21.1;q11.2) translocation. *Hum Genet* **131**, 121-130, doi:10.1007/s00439-011-1051-4 (2012).
- 159 Sharp, A. J., Spotswood, H. T., Robinson, D. O., Turner, B. M. & Jacobs, P. A. Molecular and cytogenetic analysis of the spreading of X inactivation in X;autosome translocations. *Hum Mol Genet* **11**, 3145-3156, doi:10.1093/hmg/11.25.3145 (2002).
- 160 White, W. M., Willard, H. F., Van Dyke, D. L. & Wolff, D. J. The spreading of X inactivation into autosomal material of an x;autosome translocation: evidence for a difference between autosomal and X-chromosomal DNA. *Am J Hum Genet* **63**, 20-28, doi:10.1086/301922 (1998).
- 161 Yeung, K. S. *et al.* Spread of X inactivation on chromosome 15 is associated with a more severe phenotype in a girl with an unbalanced t(X; 15) translocation. *Am J Med Genet A* **164A**, 2521-2528, doi:10.1002/ajmg.a.36670 (2014).
- 162 Cotton, A. M. *et al.* Spread of X-chromosome inactivation into autosomal sequences: role for DNA elements, chromatin features and chromosomal domains. *Hum Mol Genet* **23**, 1211-1223, doi:10.1093/hmg/ddt513 (2014).
- 163 Vassiliou, G. S. *et al.* An acquired translocation in JAK2 Val617Phe-negative essential thrombocythemia associated with autosomal spread of X-inactivation. *Haematologica* **91**, 1100-1104 (2006).
- 164 Calabrese, C. *et al.* Genomic basis for RNA alterations revealed by whole-genome analyses of 27 cancer types. *bioRxiv*, 183889, doi:10.1101/183889 (2018).
- 165 Whalley, J. P. *et al.* Framework for quality assessment of whole genome, cancer sequences. *bioRxiv*, 140921, doi:10.1101/140921 (2017).

- 166 Wala, J. A. *et al.* Selective and mechanistic sources of recurrent rearrangements across the cancer genome. *bioRxiv*, 187609, doi:10.1101/187609 (2017).
- 167 Waszak, S. M. *et al.* Germline determinants of the somatic mutation landscape in 2,642 cancer genomes. *bioRxiv*, 208330, doi:10.1101/208330 (2017).
- 168 R package 'sm': nonparametric smoothing methods v. 2.2-5.4 (2014).
- 169 Delaneau, O., Zagury, J. F. & Marchini, J. Improved whole-chromosome phasing for disease and population genetic studies. *Nat Methods* **10**, 5-6, doi:10.1038/nmeth.2307 (2013).
- 170 Delaneau, O., Howie, B., Cox, A. J., Zagury, J. F. & Marchini, J. Haplotype estimation using sequencing reads. *Am J Hum Genet* **93**, 687-696, doi:10.1016/j.ajhg.2013.09.002 (2013).
- 171 De Gregori, M. *et al.* Cryptic deletions are a common finding in "balanced" reciprocal and complex chromosome rearrangements: a study of 59 patients. *J Med Genet* **44**, 750-762, doi:10.1136/jmg.2007.052787 (2007).
- 172 R: A Language and Environment for Statistical Computing v. Version 3.5.0 (2018).
- 173 Heard, E. & Disteché, C. M. Dosage compensation in mammals: fine-tuning the expression of the X chromosome. *Genes Dev* **20**, 1848-1867, doi:10.1101/gad.1422906 (2006).
- 174 Dean, F. B., Nelson, J. R., Giesler, T. L. & Lasken, R. S. Rapid amplification of plasmid and phage DNA using Phi 29 DNA polymerase and multiply-primed rolling circle amplification. *Genome Res* **11**, 1095-1099, doi:10.1101/gr.180501 (2001).
- 175 Dean, F. B. *et al.* Comprehensive human genome amplification using multiple displacement amplification. *Proc Natl Acad Sci U S A* **99**, 5261-5266, doi:10.1073/pnas.082089499 (2002).
- 176 Gawad, C., Koh, W. & Quake, S. R. Single-cell genome sequencing: current state of the science. *Nat Rev Genet* **17**, 175-188, doi:10.1038/nrg.2015.16 (2016).
- 177 Macaulay, I. C. *et al.* G&T-seq: parallel sequencing of single-cell genomes and transcriptomes. *Nat Methods* **12**, 519-522, doi:10.1038/nmeth.3370 (2015).
- 178 Macaulay, I. C. *et al.* Separation and parallel sequencing of the genomes and transcriptomes of single cells using G&T-seq. *Nat Protoc* **11**, 2081-2103, doi:10.1038/nprot.2016.138 (2016).
- 179 Tang, F. *et al.* mRNA-Seq whole-transcriptome analysis of a single cell. *Nat Methods* **6**, 377-382, doi:10.1038/nmeth.1315 (2009).
- 180 Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* **5**, 621-628, doi:10.1038/nmeth.1226 (2008).
- 181 Nagalakshmi, U. *et al.* The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* **320**, 1344-1349, doi:10.1126/science.1158441 (2008).
- 182 Hwang, B., Lee, J. H. & Bang, D. Single-cell RNA sequencing technologies and bioinformatics pipelines. *Exp Mol Med* **50**, 96, doi:10.1038/s12276-018-0071-8 (2018).
- 183 Choi, Y. H. & Kim, J. K. Dissecting Cellular Heterogeneity Using Single-Cell RNA Sequencing. *Mol Cells* **42**, 189-199, doi:10.14348/molcells.2019.2446 (2019).

- 184 Gierahn, T. M. *et al.* Seq-Well: portable, low-cost RNA sequencing of single
cells at high throughput. *Nat Methods* **14**, 395-398, doi:10.1038/nmeth.4179
(2017).
- 185 Rosenberg, A. B. *et al.* Single-cell profiling of the developing mouse brain and
spinal cord with split-pool barcoding. *Science* **360**, 176-182,
doi:10.1126/science.aam8999 (2018).
- 186 Regev, A. *et al.* The Human Cell Atlas. *Elife* **6**, doi:10.7554/eLife.27041 (2017).
- 187 Papalexi, E. & Satija, R. Single-cell RNA sequencing to explore immune cell
heterogeneity. *Nat Rev Immunol* **18**, 35-45, doi:10.1038/nri.2017.76 (2018).
- 188 Stegle, O., Teichmann, S. A. & Marioni, J. C. Computational and analytical
challenges in single-cell transcriptomics. *Nat Rev Genet* **16**, 133-145,
doi:10.1038/nrg3833 (2015).
- 189 Zappia, L., Phipson, B. & Oshlack, A. Exploring the single-cell RNA-seq
analysis landscape with the scRNA-tools database. *PLoS Comput Biol* **14**,
e1006245, doi:10.1371/journal.pcbi.1006245 (2018).
- 190 Chen, G., Ning, B. & Shi, T. Single-Cell RNA-Seq Technologies and Related
Computational Data Analysis. *Front Genet* **10**, 317,
doi:10.3389/fgene.2019.00317 (2019).
- 191 Linnarsson, S. & Teichmann, S. A. Single-cell genomics: coming of age.
Genome Biol **17**, 97, doi:10.1186/s13059-016-0960-x (2016).
- 192 Navin, N. *et al.* Tumour evolution inferred by single-cell sequencing. *Nature*
472, 90-94, doi:10.1038/nature09807 (2011).
- 193 Navin, N. E. The first five years of single-cell cancer genomics and beyond.
Genome Res **25**, 1499-1507, doi:10.1101/gr.191098.115 (2015).
- 194 de Bourcy, C. F. *et al.* A quantitative comparison of single-cell whole genome
amplification methods. *PLoS One* **9**, e105585,
doi:10.1371/journal.pone.0105585 (2014).
- 195 Ning, L. *et al.* Quantitative assessment of single-cell whole genome
amplification methods for detecting copy number variation using hippocampal
neurons. *Sci Rep* **5**, 11415, doi:10.1038/srep11415 (2015).
- 196 Li, N. *et al.* The Performance of Whole Genome Amplification Methods and
Next-Generation Sequencing for Pre-Implantation Genetic Diagnosis of
Chromosomal Abnormalities. *J Genet Genomics* **42**, 151-159,
doi:10.1016/j.jgg.2015.03.001 (2015).
- 197 Biezuner, T. *et al.* Comparison of seven single cell Whole Genome
Amplification commercial kits using targeted sequencing. *bioRxiv*, 186940,
doi:10.1101/186940 (2017).
- 198 Borgstrom, E., Paterlini, M., Mold, J. E., Frisen, J. & Lundeberg, J. Comparison
of whole genome amplification techniques for human single cell exome
sequencing. *PLoS One* **12**, e0171566, doi:10.1371/journal.pone.0171566
(2017).
- 199 Deleye, L. *et al.* Performance of four modern whole genome amplification
methods for copy number variant detection in single cells. *Sci Rep* **7**, 3422,
doi:10.1038/s41598-017-03711-y (2017).
- 200 Mendez, P., Fang, L. T., Jablons, D. M. & Kim, I. J. Systematic comparison of
two whole-genome amplification methods for targeted next-generation
sequencing using frozen and FFPE normal and cancer tissues. *Sci Rep* **7**,
4055, doi:10.1038/s41598-017-04419-9 (2017).

- 201 Telenius, H. *et al.* Degenerate oligonucleotide-primed PCR: general
amplification of target DNA by a single degenerate primer. *Genomics* **13**, 718-
725 (1992).
- 202 Zong, C., Lu, S., Chapman, A. R. & Xie, X. S. Genome-wide detection of single-
nucleotide and copy-number variations of a single human cell. *Science* **338**,
1622-1626, doi:10.1126/science.1229164 (2012).
- 203 Langmore, J. P. Rubicon Genomics, Inc. *Pharmacogenomics* **3**, 557-560,
doi:10.1517/14622416.3.4.557 (2002).
- 204 Lasken, R. S. Genomic DNA amplification by the multiple displacement
amplification (MDA) method. *Biochem Soc Trans* **37**, 450-453,
doi:10.1042/BST0370450 (2009).
- 205 Wang, Y. *et al.* Clonal evolution in breast cancer revealed by single nucleus
genome sequencing. *Nature* **512**, 155-160, doi:10.1038/nature13600 (2014).
- 206 Bohrsen, C. L. *et al.* Linked-read analysis identifies mutations in single-cell
DNA-sequencing data. *Nat Genet* **51**, 749-754, doi:10.1038/s41588-019-0366-
2 (2019).
- 207 Wang, J., Fan, H. C., Behr, B. & Quake, S. R. Genome-wide single-cell analysis
of recombination activity and de novo mutation rates in human sperm. *Cell* **150**,
402-412, doi:10.1016/j.cell.2012.06.030 (2012).
- 208 Zafar, H., Wang, Y., Nakhleh, L., Navin, N. & Chen, K. Monovar: single-
nucleotide variant detection in single cells. *Nat Methods* **13**, 505-507,
doi:10.1038/nmeth.3835 (2016).
- 209 Edrisi, M., Zafar, H. & Nakhleh, L. A Combinatorial Approach for Single-cell
Variant Detection via Phylogenetic Inference. *bioRxiv*, 693960,
doi:10.1101/693960 (2019).
- 210 Zafar, H., Navin, N., Chen, K. & Nakhleh, L. SiCloneFit: Bayesian inference of
population structure, genotype, and phylogeny of tumor clones from single-cell
genome sequencing data. *bioRxiv*, 394262, doi:10.1101/394262 (2018).
- 211 Singer, J., Kuipers, J., Jahn, K. & Beerewinkel, N. Single-cell mutation
identification via phylogenetic inference. *Nat Commun* **9**, 5144,
doi:10.1038/s41467-018-07627-7 (2018).
- 212 Chappell, L., Russell, A. J. C. & Voet, T. Single-Cell (Multi)omics Technologies.
Annu Rev Genomics Hum Genet **19**, 15-41, doi:10.1146/annurev-genom-
091416-035324 (2018).
- 213 Hu, Y. *et al.* Single Cell Multi-Omics Technology: Methodology and Application.
Front Cell Dev Biol **6**, 28, doi:10.3389/fcell.2018.00028 (2018).
- 214 Dey, S. S., Kester, L., Spanjaard, B., Bienko, M. & van Oudenaarden, A.
Integrated genome and transcriptome sequencing of the same cell. *Nat*
Biotechnol **33**, 285-289, doi:10.1038/nbt.3129 (2015).
- 215 Han, K. Y. *et al.* SIDR: simultaneous isolation and parallel sequencing of
genomic DNA and total RNA from single cells. *Genome Res* **28**, 75-87,
doi:10.1101/gr.223263.117 (2018).
- 216 Ahn, S. J., Costa, J. & Emanuel, J. R. PicoGreen quantitation of DNA: effective
evaluation of samples pre- or post-PCR. *Nucleic Acids Res* **24**, 2623-2625,
doi:10.1093/nar/24.13.2623 (1996).
- 217 Shen, Z. *et al.* MPprimer: a program for reliable multiplex PCR primer design.
BMC Bioinformatics **11**, 143, doi:10.1186/1471-2105-11-143 (2010).
- 218 Li, H. A statistical framework for SNP calling, mutation discovery, association
mapping and population genetical parameter estimation from sequencing data.
Bioinformatics **27**, 2987-2993, doi:10.1093/bioinformatics/btr509 (2011).

- 219 Kozarewa, I. & Turner, D. J. Amplification-free library preparation for paired-end Illumina sequencing. *Methods Mol Biol* **733**, 257-266, doi:10.1007/978-1-61779-089-8_18 (2011).
- 220 Tischler, G. & Leonard, S. biobambam: tools for read pair collation based algorithms on BAM files. *Source Code Biol Med* **9**, doi:10.1186/1751-0473-9-13 (2014).
- 221 Martin, M. Cutadapt Removes Adapter Sequences From High-Throughput Sequencing Reads. *EMBnet.journal*, doi:<https://doi.org/10.14806/ej.17.1.200>. (2011).
- 222 Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv:1303.3997* (2013).
- 223 Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079, doi:10.1093/bioinformatics/btp352 (2009).
- 224 BroadInstitute. <http://broadinstitute.github.io/picard/>, (2012).
- 225 ineq: Measuring Inequality, Concentration, and Poverty. v. R package version 0.2-13 (2014).
- 226 MESS: Miscellaneous Esoteric Statistical Scripts. v. R package version 0.5.5 (2019).
- 227 Sherman, M. A. *et al.* PaSD-qc: quality control for single cell whole-genome sequencing data using power spectral density estimation. *Nucleic Acids Res* **46**, e20, doi:10.1093/nar/gkx1195 (2018).
- 228 Picher, A. J. *et al.* TruePrime is a novel method for whole-genome amplification from single cells based on TthPrimPol. *Nat Commun* **7**, 13296, doi:10.1038/ncomms13296 (2016).
- 229 Leung, M. L., Wang, Y., Waters, J. & Navin, N. E. SNES: single nucleus exome sequencing. *Genome Biol* **16**, 55, doi:10.1186/s13059-015-0616-2 (2015).
- 230 Tau, G. Z. & Peterson, B. S. Normal development of brain circuits. *Neuropsychopharmacology* **35**, 147-168, doi:10.1038/npp.2009.115 (2010).
- 231 Herculano-Houzel, S. The human brain in numbers: a linearly scaled-up primate brain. *Front Hum Neurosci* **3**, 31, doi:10.3389/neuro.09.031.2009 (2009).
- 232 Marin, O. & Muller, U. Lineage origins of GABAergic versus glutamatergic neurons in the neocortex. *Curr Opin Neurobiol* **26**, 132-141, doi:10.1016/j.conb.2014.01.015 (2014).
- 233 Kriegstein, A. R. Constructing circuits: neurogenesis and migration in the developing neocortex. *Epilepsia* **46 Suppl 7**, 15-21, doi:10.1111/j.1528-1167.2005.00304.x (2005).
- 234 Jiang, X. & Nardelli, J. Cellular and molecular introduction to brain development. *Neurobiol Dis* **92**, 3-17, doi:10.1016/j.nbd.2015.07.007 (2016).
- 235 Duque, A. & Spector, R. A balanced evaluation of the evidence for adult neurogenesis in humans: implication for neuropsychiatric disorders. *Brain Struct Funct*, doi:10.1007/s00429-019-01917-6 (2019).
- 236 Malik, S. *et al.* Neurogenesis continues in the third trimester of pregnancy and is suppressed by premature birth. *J Neurosci* **33**, 411-423, doi:10.1523/JNEUROSCI.4445-12.2013 (2013).
- 237 Evrony, G. D. One brain, many genomes. *Science* **354**, 557-558, doi:10.1126/science.aak9761 (2016).
- 238 Yurov, Y. B. *et al.* Aneuploidy and confined chromosomal mosaicism in the developing human brain. *PLoS One* **2**, e558, doi:10.1371/journal.pone.0000558 (2007).

- 239 Chronister, W. D. *et al.* Neurons with Complex Karyotypes Are Rare in Aged Human Neocortex. *Cell Rep* **26**, 825-835 e827, doi:10.1016/j.celrep.2018.12.107 (2019).
- 240 Muotri, A. R. *et al.* Somatic mosaicism in neuronal precursor cells mediated by L1 retrotransposition. *Nature* **435**, 903-910, doi:10.1038/nature03663 (2005).
- 241 Coufal, N. G. *et al.* L1 retrotransposition in human neural progenitor cells. *Nature* **460**, 1127-1131, doi:10.1038/nature08248 (2009).
- 242 Baillie, J. K. *et al.* Somatic retrotransposition alters the genetic landscape of the human brain. *Nature* **479**, 534-537, doi:10.1038/nature10531 (2011).
- 243 Knouse, K. A., Wu, J. & Hendricks, A. Detection of Copy Number Alterations Using Single Cell Sequencing. *J Vis Exp*, doi:10.3791/55143 (2017).
- 244 Pamphlett, R. Somatic mutation: a cause of sporadic neurodegenerative diseases? *Med Hypotheses* **62**, 679-682, doi:10.1016/j.mehy.2003.11.023 (2004).
- 245 Frank, S. A. Evolution in health and medicine Sackler colloquium: Somatic evolutionary genomics: mutations during development cause highly variable genetic mosaicism with risk of cancer and neurodegeneration. *Proc Natl Acad Sci U S A* **107 Suppl 1**, 1725-1730, doi:10.1073/pnas.0909343106 (2010).
- 246 Verheijen, B. M., Vermulst, M. & van Leeuwen, F. W. Somatic mutations in neurons during aging and neurodegeneration. *Acta Neuropathol* **135**, 811-826, doi:10.1007/s00401-018-1850-y (2018).
- 247 Zhao, B. *et al.* Somatic LINE-1 retrotransposition in cortical neurons and non-brain tissues of Rett patients and healthy individuals. *PLoS Genet* **15**, e1008043, doi:10.1371/journal.pgen.1008043 (2019).
- 248 Mullen, R. J., Buck, C. R. & Smith, A. M. NeuN, a neuronal specific nuclear protein in vertebrates. *Development* **116**, 201-211 (1992).
- 249 Baker, S. C. *et al.* The External RNA Controls Consortium: a progress report. *Nat Methods* **2**, 731-734, doi:10.1038/nmeth1005-731 (2005).
- 250 Picelli, S. *et al.* Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nat Methods* **10**, 1096-1098, doi:10.1038/nmeth.2639 (2013).
- 251 Picelli, S. *et al.* Full-length RNA-seq from single cells using Smart-seq2. *Nat Protoc* **9**, 171-181, doi:10.1038/nprot.2014.006 (2014).
- 252 Syed, F. Application of Nextera™ technology to RNA-seq library preparation. *Nature Methods* **7**, 1026, doi:<https://doi.org/10.1038/nmeth.f.317> (2010).
- 253 Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21, doi:10.1093/bioinformatics/bts635 (2013).
- 254 Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923-930, doi:10.1093/bioinformatics/btt656 (2014).
- 255 McCarthy, D. J., Campbell, K. R., Lun, A. T. & Wills, Q. F. Scater: pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R. *Bioinformatics* **33**, 1179-1186, doi:10.1093/bioinformatics/btw777 (2017).
- 256 Wang, L., Wang, S. & Li, W. RSeQC: quality control of RNA-seq experiments. *Bioinformatics* **28**, 2184-2185, doi:10.1093/bioinformatics/bts356 (2012).
- 257 Kiselev, V. Y. *et al.* SC3: consensus clustering of single-cell RNA-seq data. *Nat Methods* **14**, 483-486, doi:10.1038/nmeth.4236 (2017).
- 258 McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**, 1297-1303, doi:10.1101/gr.107524.110 (2010).

- 259 Jones, D. *et al.* cgpCaVEManWrapper: Simple Execution of CaVEMan in Order
to Detect Somatic Single Nucleotide Variants in NGS Data. *Curr Protoc*
Bioinformatics **56**, 15 10 11-15 10 18, doi:10.1002/cpbi.20 (2016).
- 260 Haeussler, M. *et al.* The UCSC Genome Browser database: 2019 update.
Nucleic Acids Res **47**, D853-D858, doi:10.1093/nar/gky1095 (2019).
- 261 Jahn, K., Kuipers, J. & Beerenwinkel, N. Tree inference for single-cell data.
Genome Biol **17**, 86, doi:10.1186/s13059-016-0936-x (2016).
- 262 Paradis, E. & Schliep, K. ape 5.0: an environment for modern phylogenetics
and evolutionary analyses in R. *Bioinformatics* **35**, 526-528,
doi:10.1093/bioinformatics/bty633 (2019).
- 263 Buels, R. *et al.* JBrowse: a dynamic web platform for genome visualization and
analysis. *Genome Biol* **17**, 66, doi:10.1186/s13059-016-0924-1 (2016).
- 264 Akalin, A., Franke, V., Vlahovicek, K., Mason, C. E. & Schubeler, D.
Genomation: a toolkit to summarize, annotate and visualize genomic intervals.
Bioinformatics **31**, 1127-1129, doi:10.1093/bioinformatics/btu775 (2015).
- 265 McLean, C. Y. *et al.* GREAT improves functional interpretation of cis-regulatory
regions. *Nat Biotechnol* **28**, 495-501, doi:10.1038/nbt.1630 (2010).
- 266 rGREAT: Client for GREAT Analysis. <https://github.com/jokergoo/rGREAT>
(2019).
- 267 Lawrence, M. S. *et al.* Mutational heterogeneity in cancer and the search for
new cancer-associated genes. *Nature* **499**, 214-218, doi:10.1038/nature12213
(2013).
- 268 Kodama, T. *et al.* Neuronal classification and marker gene identification via
single-cell expression profiling of brainstem vestibular neurons subserving
cerebellar learning. *J Neurosci* **32**, 7819-7831, doi:10.1523/JNEUROSCI.0543-
12.2012 (2012).
- 269 Darmanis, S. *et al.* A survey of human brain transcriptome diversity at the single
cell level. *Proc Natl Acad Sci U S A* **112**, 7285-7290,
doi:10.1073/pnas.1507125112 (2015).
- 270 Lake, B. B. *et al.* Neuronal subtypes and diversity revealed by single-nucleus
RNA sequencing of the human brain. *Science* **352**, 1586-1590,
doi:10.1126/science.aaf1204 (2016).
- 271 Habib, N. *et al.* Massively parallel single-nucleus RNA-seq with DroNc-seq. *Nat*
Methods **14**, 955-958, doi:10.1038/nmeth.4407 (2017).
- 272 Yeh, F. L. *et al.* SV2 mediates entry of tetanus neurotoxin into central neurons.
PLoS Pathog **6**, e1001207, doi:10.1371/journal.ppat.1001207 (2010).
- 273 Nishikura, K. A-to-I editing of coding and non-coding RNAs by ADARs. *Nat Rev*
Mol Cell Biol **17**, 83-96, doi:10.1038/nrm.2015.4 (2016).
- 274 Franco, I. *et al.* Somatic mutagenesis in satellite cells associates with human
skeletal muscle aging. *Nat Commun* **9**, 800, doi:10.1038/s41467-018-03244-6
(2018).
- 275 Chen, C. *et al.* Single-cell whole-genome analyses by Linear Amplification via
Transposon Insertion (LIANTI). *Science* **356**, 189-194,
doi:10.1126/science.aak9787 (2017).
- 276 Marinov, G. K. *et al.* From single-cell to cell-pool transcriptomes: stochasticity
in gene expression and RNA splicing. *Genome Res* **24**, 496-510,
doi:10.1101/gr.161034.113 (2014).
- 277 Deng, Q., Ramskold, D., Reinius, B. & Sandberg, R. Single-cell RNA-seq
reveals dynamic, random monoallelic gene expression in mammalian cells.
Science **343**, 193-196, doi:10.1126/science.1245316 (2014).

- 278 Borel, C. *et al.* Biased allelic expression in human primary fibroblast single cells. *Am J Hum Genet* **96**, 70-80, doi:10.1016/j.ajhg.2014.12.001 (2015).
- 279 Reinius, B. *et al.* Analysis of allelic expression patterns in clonal somatic cells by single-cell RNA-seq. *Nat Genet* **48**, 1430-1435, doi:10.1038/ng.3678 (2016).
- 280 Lake, B. B. *et al.* A comparative strategy for single-nucleus and single-cell transcriptomes confirms accuracy in predicted cell-type expression from nuclear RNA. *Sci Rep* **7**, 6031, doi:10.1038/s41598-017-04426-w (2017).
- 281 Bakken, T. E. *et al.* Single-nucleus and single-cell transcriptomes compared in matched cortical cell types. *PLoS One* **13**, e0209648, doi:10.1371/journal.pone.0209648 (2018).
- 282 Bazak, L. *et al.* A-to-I RNA editing occurs at over a hundred million genomic sites, located in a majority of human genes. *Genome Res* **24**, 365-376, doi:10.1101/gr.164749.113 (2014).
- 283 Picardi, E. *et al.* Profiling RNA editing in human tissues: towards the inosinome Atlas. *Sci Rep* **5**, 14941, doi:10.1038/srep14941 (2015).
- 284 Brusa, R. *et al.* Early-onset epilepsy and postnatal lethality associated with an editing-deficient GluR-B allele in mice. *Science* **270**, 1677-1680, doi:10.1126/science.270.5242.1677 (1995).
- 285 Picardi, E., Horner, D. S. & Pesole, G. Single-cell transcriptomics reveals specific RNA editing signatures in the human brain. *RNA* **23**, 860-865, doi:10.1261/rna.058271.116 (2017).
- 286 Xu, X. *et al.* Single-cell exome sequencing reveals single-nucleotide mutation characteristics of a kidney tumor. *Cell* **148**, 886-895, doi:10.1016/j.cell.2012.02.025 (2012).
- 287 Petljak, M. *et al.* Characterizing Mutational Signatures in Human Cancer Cell Lines Reveals Episodic APOBEC Mutagenesis. *Cell* **176**, 1282-1294 e1220, doi:10.1016/j.cell.2019.02.012 (2019).
- 288 Ziegenhain, C. *et al.* Comparative Analysis of Single-Cell RNA Sequencing Methods. *Mol Cell* **65**, 631-643 e634, doi:10.1016/j.molcel.2017.01.023 (2017).
- 289 Ding, J. *et al.* Systematic comparative analysis of single cell RNA-sequencing methods. *bioRxiv*, 632216, doi:10.1101/632216 (2019).
- 290 Toivanen, R. & Shen, M. M. Prostate organogenesis: tissue induction, hormonal regulation and cell type specification. *Development* **144**, 1382-1398, doi:10.1242/dev.148270 (2017).
- 291 McNeal, J. E. The zonal anatomy of the prostate. *Prostate* **2**, 35-49, doi:10.1002/pros.2990020105 (1981).
- 292 Lee, C. H., Akin-Olugbade, O. & Kirschenbaum, A. Overview of prostate anatomy, histology, and pathology. *Endocrinol Metab Clin North Am* **40**, 565-575, viii-ix, doi:10.1016/j.ecl.2011.05.012 (2011).
- 293 Aaron, L., Franco, O. E. & Hayward, S. W. Review of Prostate Anatomy and Embryology and the Etiology of Benign Prostatic Hyperplasia. *Urol Clin North Am* **43**, 279-288, doi:10.1016/j.ucl.2016.04.012 (2016).
- 294 McNeal, J. E. Regional morphology and pathology of the prostate. *Am J Clin Pathol* **49**, 347-357, doi:10.1093/ajcp/49.3.347 (1968).
- 295 Moad, M. *et al.* Multipotent Basal Stem Cells, Maintained in Localized Proximal Niches, Support Directed Long-Ranging Epithelial Flows in Human Prostates. *Cell Rep* **20**, 1609-1622, doi:10.1016/j.celrep.2017.07.061 (2017).
- 296 El-Alfy, M., Pelletier, G., Hermo, L. S. & Labrie, F. Unique features of the basal cells of human prostate epithelium. *Microsc Res Tech* **51**, 436-446, doi:10.1002/1097-0029(20001201)51:5<436::AID-JEMT6>3.0.CO;2-T (2000).

- 297 Shen, M. M. & Abate-Shen, C. Molecular genetics of prostate cancer: new prospects for old challenges. *Genes Dev* **24**, 1967-2000, doi:10.1101/gad.1965810 (2010).
- 298 Henry, G. H. *et al.* A Cellular Anatomy of the Normal Adult Human Prostate and Prostatic Urethra. *Cell Rep* **25**, 3530-3542 e3535, doi:10.1016/j.celrep.2018.11.086 (2018).
- 299 Wang, Y., Hayward, S., Cao, M., Thayer, K. & Cunha, G. Cell differentiation lineage in the prostate. *Differentiation* **68**, 270-279 (2001).
- 300 Xue, Y. *et al.* Proliferative activity and branching morphogenesis in the human prostate: a closer look at pre- and postnatal prostate growth. *Prostate* **49**, 132-139, doi:10.1002/pros.1127 (2001).
- 301 Cunha, G. R. *et al.* Development of the human prostate. *Differentiation* **103**, 24-45, doi:10.1016/j.diff.2018.08.005 (2018).
- 302 Hannezo, E. *et al.* A Unifying Theory of Branching Morphogenesis. *Cell* **171**, 242-255 e227, doi:10.1016/j.cell.2017.08.026 (2017).
- 303 Cunha, G. R. *et al.* The endocrinology and developmental biology of the prostate. *Endocr Rev* **8**, 338-362, doi:10.1210/edrv-8-3-338 (1987).
- 304 Dhanasekaran, S. M. *et al.* Molecular profiling of human prostate tissues: insights into gene expression patterns of prostate development during puberty. *FASEB J* **19**, 243-245, doi:10.1096/fj.04-2415fje (2005).
- 305 Zaichick, S. & Zaichick, V. Relations of morphometric parameters to zinc content in paediatric and nonhyperplastic young adult prostate glands. *Andrology* **1**, 139-146, doi:10.1111/j.2047-2927.2012.00005.x (2013).
- 306 Roehrborn, C. G. Benign prostatic hyperplasia: an overview. *Rev Urol* **7 Suppl 9**, S3-S14 (2005).
- 307 Attard, G. *et al.* Prostate cancer. *Lancet* **387**, 70-82, doi:10.1016/S0140-6736(14)61947-4 (2016).
- 308 Chokkalingam, A. P. *et al.* Prostate carcinoma risk subsequent to diagnosis of benign prostatic hyperplasia: a population-based cohort study in Sweden. *Cancer* **98**, 1727-1734, doi:10.1002/cncr.11710 (2003).
- 309 Dai, X., Fang, X., Ma, Y. & Xianyu, J. Benign Prostatic Hyperplasia and the Risk of Prostate Cancer and Bladder Cancer: A Meta-Analysis of Observational Studies. *Medicine (Baltimore)* **95**, e3493, doi:10.1097/MD.00000000000003493 (2016).
- 310 Schenk, J. M. *et al.* Association of symptomatic benign prostatic hyperplasia and prostate cancer: results from the prostate cancer prevention trial. *Am J Epidemiol* **173**, 1419-1428, doi:10.1093/aje/kwq493 (2011).
- 311 Humphrey, P. A. Histological variants of prostatic carcinoma and their significance. *Histopathology* **60**, 59-74, doi:10.1111/j.1365-2559.2011.04039.x (2012).
- 312 Cheng, L. *et al.* Evidence of independent origin of multiple tumors from patients with prostate cancer. *J Natl Cancer Inst* **90**, 233-237, doi:10.1093/jnci/90.3.233 (1998).
- 313 Lindberg, J. *et al.* Exome sequencing of prostate cancer supports the hypothesis of independent tumour origins. *Eur Urol* **63**, 347-353, doi:10.1016/j.eururo.2012.03.050 (2013).
- 314 Boutros, P. C. *et al.* Spatial genomic heterogeneity within localized, multifocal prostate cancer. *Nat Genet* **47**, 736-745, doi:10.1038/ng.3315 (2015).
- 315 Cooper, C. S. *et al.* Analysis of the genetic phylogeny of multifocal prostate cancer identifies multiple independent clonal expansions in neoplastic and

- morphologically normal prostate tissue. *Nat Genet* **47**, 367-372, doi:10.1038/ng.3221 (2015).
- 316 Boyd, L. K. *et al.* High-resolution genome-wide copy-number analysis suggests a monoclonal origin of multifocal prostate cancer. *Genes Chromosomes Cancer* **51**, 579-589, doi:10.1002/gcc.21944 (2012).
- 317 Armenia, J. *et al.* The long tail of oncogenic drivers in prostate cancer. *Nat Genet* **50**, 645-651, doi:10.1038/s41588-018-0078-z (2018).
- 318 Tomlins, S. A. *et al.* Distinct classes of chromosomal rearrangements create oncogenic ETS gene fusions in prostate cancer. *Nature* **448**, 595-599, doi:10.1038/nature06024 (2007).
- 319 Gasi Tandefelt, D., Boormans, J., Hermans, K. & Trapman, J. ETS fusion genes in prostate cancer. *Endocr Relat Cancer* **21**, R143-152, doi:10.1530/ERC-13-0390 (2014).
- 320 Gundem, G. *et al.* The evolutionary history of lethal metastatic prostate cancer. *Nature* **520**, 353-357, doi:10.1038/nature14347 (2015).
- 321 Wedge, D. C. *et al.* Sequencing of prostate cancers identifies new cancer genes, routes of progression and drug targets. *Nat Genet* **50**, 682-692, doi:10.1038/s41588-018-0086-z (2018).
- 322 Montagnani, S., Rueger, M. A., Hosoda, T. & Nurzynska, D. Adult Stem Cells in Tissue Maintenance and Regeneration. *Stem Cells Int* **2016**, 7362879, doi:10.1155/2016/7362879 (2016).
- 323 Jaworska, D., Krol, W. & Szliszka, E. Prostate Cancer Stem Cells: Research Advances. *Int J Mol Sci* **16**, 27433-27449, doi:10.3390/ijms161126036 (2015).
- 324 Richardson, G. D. *et al.* CD133, a novel marker for human prostatic epithelial stem cells. *J Cell Sci* **117**, 3539-3545, doi:10.1242/jcs.01222 (2004).
- 325 Hudson, D. L., O'Hare, M., Watt, F. M. & Masters, J. R. Proliferative heterogeneity in the human prostate: evidence for epithelial stem cells. *Lab Invest* **80**, 1243-1250 (2000).
- 326 Collins, A. T., Habib, F. K., Maitland, N. J. & Neal, D. E. Identification and isolation of human prostate epithelial stem cells based on alpha(2)beta(1)-integrin expression. *J Cell Sci* **114**, 3865-3872 (2001).
- 327 Goldstein, A. S. *et al.* Trop2 identifies a subpopulation of murine and human prostate basal cells with stem cell characteristics. *Proc Natl Acad Sci U S A* **105**, 20882-20887, doi:10.1073/pnas.0811411106 (2008).
- 328 Leong, K. G., Wang, B. E., Johnson, L. & Gao, W. Q. Generation of a prostate from a single adult stem cell. *Nature* **456**, 804-808, doi:10.1038/nature07427 (2008).
- 329 Karthaus, W. R. *et al.* Identification of multipotent luminal progenitor cells in human prostate organoid cultures. *Cell* **159**, 163-175, doi:10.1016/j.cell.2014.08.017 (2014).
- 330 Zhang, D. *et al.* Stem cell and neurogenic gene-expression profiles link prostate basal cells to aggressive prostate cancer. *Nat Commun* **7**, 10798, doi:10.1038/ncomms10798 (2016).
- 331 Gaisa, N. T. *et al.* Clonal architecture of human prostatic epithelium in benign and malignant conditions. *J Pathol* **225**, 172-180, doi:10.1002/path.2959 (2011).
- 332 Blackwood, J. K. *et al.* In situ lineage tracking of human prostatic epithelial stem cell fate reveals a common clonal origin for basal and luminal cells. *J Pathol* **225**, 181-188, doi:10.1002/path.2965 (2011).

- 333 Fiala, J. C. Reconstruct: a free editor for serial section microscopy. *J Microsc*
 218, 52-61, doi:10.1111/j.1365-2818.2005.01466.x (2005).
- 334 plotly for R (2018).
- 335 BRASS - Breakpoints via assembly v. 6.2.0 (2019).
- 336 Kuznetsova, A., Brockhoff, P. & Christensen, R. lmerTest Package: Tests in
 Linear Mixed Effects Models. *Journal of Statistical Software* **82**, 1-26,
 doi:10.18637/jss.v082.i13 (2017).
- 337 Fox, J. & Weisberg, S. Visualizing Fit and Lack of Fit in Complex Regression
 Models with Predictor Effect Plots and Partial Residuals. *Journal of Statistical
 Software* **87**, 1-27 (2018).
- 338 dndscv: Poisson-based dN/dS models to quantify natural selection in somatic
 evolution v. R package version 0.0.1.0 (2019).
- 339 Gonzalez-Perez, A. *et al.* IntOGen-mutations identifies cancer drivers across
 tumor types. *Nat Methods* **10**, 1081-1082, doi:10.1038/nmeth.2642 (2013).
- 340 Farmery, J. H. R., Smith, M. L., Diseases, N. B.-R. & Lynch, A. G. Telomerecat:
 A ploidy-agnostic method for estimating telomere length from whole genome
 sequencing data. *Sci Rep* **8**, 1300, doi:10.1038/s41598-017-14403-y (2018).
- 341 Gerhardt, J. *et al.* FOXA1 promotes tumor progression in prostate cancer and
 represents a novel hallmark of castration-resistant prostate cancer. *Am J Pathol*
180, 848-861, doi:10.1016/j.ajpath.2011.10.021 (2012).
- 342 Cancer Genome Atlas Research, N. The Molecular Taxonomy of Primary
 Prostate Cancer. *Cell* **163**, 1011-1025, doi:10.1016/j.cell.2015.10.025 (2015).
- 343 Adams, E. J. *et al.* FOXA1 mutations alter pioneering activity, differentiation
 and prostate cancer phenotypes. *Nature* **571**, 408-412, doi:10.1038/s41586-
 019-1318-9 (2019).
- 344 Woodworth, M. B., Girsakis, K. M. & Walsh, C. A. Building a lineage from single
 cells: genetic techniques for cell lineage tracking. *Nat Rev Genet* **18**, 230-244,
 doi:10.1038/nrg.2016.159 (2017).
- 345 Yin, H., Price, F. & Rudnicki, M. A. Satellite cells and the muscle stem cell niche.
Physiol Rev **93**, 23-67, doi:10.1152/physrev.00043.2011 (2013).

List of Abbreviations and Acronyms

ASC	Adult Stem Cell
ASE	Allele-specific Expression
ASMD	Median Alignment Score
AUROC	Area Under Receiver Operating Characteristic
BLCA	Bladder Urothelial Cancer
bp	Base Pair
BPH	Benign Prostate Hyperplasia
cDNA	Copy DNA
CML	Chronic Myeloid Leukaemia
CPM	Counts per Million Mapped Reads
DHT	Dihydrotestosterone
DSB	Double-strand Break
FACS	Fluorescence-activated Cell Sorting
FISH	Fluorescence in-situ Hybridisation
GBM	Glioblastoma
gDNA	Genomic DNA
GenomiPhi	Single Cell GenomiPhi DNA Amplification Kit from GE Healthcare Life Sciences
gHet	Germline Heterozygous SNP
GLM	Generalised Linear Model
GO	Gene Ontology
H&E	Haematoxylin and Eosin Staining
HCC	Hepatocellular Carcinoma
ICGC	International Cancer Genome Consortium
IPC	Intermediate Progenitor Cell
iPSC	Induced-pluripotent Stem Cell
kb	Kilobase
LCM	Laser-Capture Microscopy
Mb	Megabase
MDA	Multiple Displacement Amplification
MDS	Multidimensional Scaling
MRCA	Most Recent Common Ancestor

mRNA	Messenger RNA
n-HDP	N-Dimensional Hierarchical Dirichlet Process
NE	Neuroendocrine Cells
NEC	Neuroepithelial Cells
OLS	Ordinary Least Squares Regression
OSC	Ovarian Serous Carcinoma
PCA	Principal Component Analysis
PCAWG	Pan-Cancer Analysis of Whole Genomes Project
QC	Quality Control
RepliG	Repli-g Single Cell Kit from Qiagen
ROS	Reactive Oxygen Species
RT	Reverse Transcription
Sanger Pipelines	Core Facility Pipelines at the Wellcome Sanger Institute
scRNA-seq	Single Cell RNA-seq
SMG	Significantly Mutated Gene
SNCT	Somatic Nuclear Cell Transfer
SNP	Single Nucleotide Polymorphism
SNV	Single Nucleotide Variant
SVZ	Subventricular Zone
TCGA	The Cancer Genome Atlas
Trueprime	Trueprime Whole Genome Amplification Kit from Expedeon
TSS	Transcription Start Site
UGS	Urogenital Sinus
VAF	Variant Allele Fraction
VZ	Ventricular Zone
WGA	Whole-genome Amplification
WGS	Whole-genome Sequencing